ON THE DESIGN OF VISUAL FEEDBACK FOR THE REHABILITATION OF HEARING-IMPAIRED SPEECH

Fabrizio Carraro

A thesis submitted for the degree of
Doctor of Philosophy

1997

Abstract

Hearing-impaired people have difficulties in developing normal language skills because of their lack of vocal feedback. Visual feedback, as a substitute for vocal feedback, has been used for many years by speech therapists in rehabilitation schemes. However clients and therapists have to cope with problems such as negative reinforcement, frustration, lack of motivation, and the high cost associated with these visual feedback approaches. This thesis analyses these problems, and describes a novel approach to designing visual feedback for the rehabilitation of hearing-impaired speech. This novel approach takes into account previous research on visual display design techniques, necessary for implementing user-friendly graphic interfaces, together with the experience and comments from speech therapists using both traditional methods and computer-aided systems, invaluable for understanding what is missing or wrong in present speech rehabilitation systems. The thesis details original experiments to investigate the best way to visually represent a specific speech feature. A novel experimental method is used where, instead of proposing different visual representations for the various speech features (such as loudness, pitch, vowel quality etc.) and assessing which work best, the various visual stimuli are shown to the subject without specifying the associated speech feature. In this way an intuitive connection between visual stimuli and speech features (the sound produced by the subject) can be characterised. The goal of the experiments is to identify the best associations for visual stimuli and speech features. The visual stimuli for each association is then used in a visual feedback scheme for that speech feature. The result of these studies are merged with real-time and system cost considerations in order to design and implement a set of modules for effective hearing-impaired speech rehabilitation. Trials results with a cohort of deaf subjects are presented for a system which includes a range of visual stimuli approaches.

Acknowledgements

This thesis allowed me to have one of the best time I had in my life. Moving to Scotland has been an immense life experience for me, and the topic of the thesis, hearing-impaired speech rehabilitation, made me know a world that gave me so much gratification and emotion. While I write these words I know that I'll go back to my home country soon. I would like to thank all the persons who contributed in making me feel at home here. Many of them will stay in my heart all my life.

First of all, I wish to express my gratitude to Prof. Mervyn Jack as supervisor of this thesis. He encouraged me to go for it, and helped me all the time in many ways, also giving me a job in his department, which solved in the best way the problem of supporting myself. Thanks to Dr. Steve Hiller, who, although acted as my second supervisor only for a short period of time, decisively helped me in designing the final structure to this thesis. A special thanks to my friend Keith Edwards, who patiently proof-read my bad writing, turning my odd English into a readable one. Furthermore, his knowledge of phonetics and linguistics gave me many helpful advise and suggestions. Another person who has been really invaluable for me is Nikki Pattison who, with her lovely and motivating attitude, helped me a lot in keep on going. Thanks to Dr. Andrew Sutherland for helping me in focusing the topic of this thesis, and Dr. Eddie Rooney, Rebecca Vaughan and Katie Robertson, for their help on the literature survey. Many thanks to Dr. Maurilio Nunes Vieira, with whom I spent many hours discussing algorithms, voice pathology, and music. Maurilio not only was able to clarify my many doubts about speech processing issues, but also gave a decisive contribution in some of the analysis modules used in the prototype system. Thanks to Dr. Mark Schmidt, who at the beginning of this thesis gave me many ideas and motivation to continue, and helped in the evaluation of the experiments, together with Dr. Fergus Mc.Innes, Dr. John Foster, Dr. Steve Love, and Keith Edwards, who I thank again. Thanks to the speech therapists Alison McDonald, Jayne Inscoe, Mhairi Gillgillian, Marion McNab, Kim Davidson-Kelly and Sarah Worsfold for giving me their time for discussing on speech rehabilitation techniques, and for finding hearing-impaired subjects and organising sessions for the experiments. A special thanks to Dr. Ian Nairn and Janette Nairn, for organising sessions for the experiments with normal hearing children, and also for inviting me for excellent dinners after that. Thanks to Dr. Alice Turk for giving me ideas and motivation and Will Dempsey and Liam Parker for hints on animations. Thanks to all the hearing-impaired persons who participated in the experiments, and from whom I've learned a lot. I hope this thesis is the beginning of the process for giving them back something in change. Finally a particular thank-you to my beloved friends and flatmates Aphrodite, Nucy and Gregorius who indirectly contributed to the successful development of this thesis by sharing the everyday life with me, surrounding me with a lively and stimulating environment.

Contents

CHAPTER 1

Introduction

This thesis describes a novel approach to designing visual feedback for the rehabilitation of hearing-impaired speech. Visual feedback provides the hearing impaired with information about their own speech that is missing because of their impairment. There have been many attempts to build effective visual feedback systems, but only a few give some limited positive results. More effective visual feedback is needed since traditional methods suffer from problems such as negative reinforcement, frustration, lack of motivation, and high cost. This thesis highlights the fact that problems are caused by a series of reasons such as: lack of good visual interface design, causing difficulties in the use of the speech rehabilitation system; scarce research in the field of visual/auditory perception, causing the design of non-intuitive visual feedback methods; relatively scarce evaluation from real users, resulting in systems which can give negative reinforcement, frustration, and ineffectiveness; and unreliable speech processing, giving insufficiently accurate results, so causing further frustration and negative reinforcement.

A novel approach is proposed, based on the following steps: 1) background knowledge of the vast literature of visual interface design; 2) location of the weak points and inconsistencies in the techniques used in traditional speech rehabilitation systems, and consideration of the techniques which were not explored; 3) selection of a set of visual stimuli as candidates for visual feedback; 4) experimentation to test the response of hearing-impaired subjects to these visual stimuli, in terms of intuitive links between visual dimensions and speech features; 5) evaluation of the experiments with the goal of using the visual stimuli that gave a good match with speech features for use as a visual feedback; 6) design, implementation and user assessment of a prototype speech rehabilitation system, with a close interaction with therapists and hearing-impaired users.

In this thesis, the steps are supported by background information on each topic. Readers who are not familiar with the speech production process are advised to read this thesis from the beginning. Readers who are familiar with the speech production process but who are not acquainted with the problems related to hearing-impaired speech may start from Section 2.3, where causes of deafness and its effects on speech are discussed. Readers who are familiar with hearing-impaired speech rehabilitation may start from Chapter 3, to be aware of the state of the art of studies in this area and published systems. The original work starts with Chapter 4, where the problems to be solved are highlighted, and a novel approach to designing visual feedback for the rehabilitation of hearing-impaired speech is proposed.

The contents of the chapters are summarised below:

Chapter 2 presents information on normal speech production, explaining how *hearing* is an essential component of the *speech chain*, and describing the production of vowels and consonants by means of the Acoustical Theory of Speech Production. This theory gives the reader the appropriate background for understanding the speech analysis algorithms used in the thesis. Causes of hearing impairment, and how this affects the normal speech production are then discussed in detail, giving a survey of the characteristics of hearing-impaired speech for different speech features, such as intensity, fundamental frequency, vowels and diphthongs, consonants, voice quality, timing, pausing, breath control, and rhythm.

Chapter 3 reviews visual feedback techniques in the rehabilitation of hearing-impaired speech, giving an extensive survey of the studies in the field, covering topics such as user groups, the physical basis of feedback, instructional context, and amount and type of information displayed as feedback. A review of published systems is also included, with many examples of screens used in the feedback for rehabilitating different speech features.

Chapter 4 discusses the problems that therapists and users encounter when using traditional visual feedback techniques, and proposes a novel approach for designing visual feedback. The problems resulted from a survey of 12 British therapists for the hearing impaired, teachers of the deaf, linguists and student speech therapists. These problems include negative reinforcement, lack of motivation, frustration, inflexibility, and difficulties in understanding how to correct speech productions. The novel approach is then described, and the first step of the novel approach is then carried-out, with an introduction to visual interface design. The introduction summarises all the most important studies in the field, covering topics such as amount of information presented, placement of information, coding of information, multimedia, and virtual reality. Some fundamentals on human visual information processing, and on feedback control mechanisms of human behaviour are also reported: the former plays an important role in the design of a visual interface, and the latter give information about the response time that a visual feedback should achieve when used in a closed loop with the user, as in the case of speech rehabilitation. The chapter ends by highlighting the need for experimentation on intuitive methods of feedback. This need is supported by a survey of literature in the field, where it appears that present visual feedback for speech rehabilitation is based on practical approaches rather than on research.

Chapter 5 covers the experiments. Three set of experiments are described. The goal of Experiment 1 is to study which features of the voice are intuitively linked to different graphic modalities. A set of appropriate visual stimuli are selected, considering a wide range of visual coding methods, including all the techniques used in actual systems, and considering new techniques. A novel experimental

method is used where, instead of proposing different visual representations for the various speech features (such as loudness, pitch, vowel quality etc.) and assessing which work best, the various visual stimuli are shown to the subject without specifying the associated speech feature. In this way an intuitive connection between visual stimuli and speech features (the sound produced by the subject) can be characterised. Experiment 2 compares two different graphic techniques for showing the same visual stimuli. The first is an animation in simulated 3D shown on a conventional computer screen, the second is the same animation in true stereo-vision, shown on a pair of 3D goggles. Experiment 3 assesses subjects' motivation when watching visual stimuli using multimedia technology.

Chapter 6 presents the design and implementation of visual feedback based on the results of the experiments, on therapists' recommendations, and on real-time considerations. For each speech feature one or more visual feedback modes is considered, in order to design a set of modules for effective hearing-impaired speech rehabilitation. Visual display guidelines are followed to achieve intuitive control of each element of the visual feedback. A help system using multimedia technology is also designed. The sections on implementation consider problems related to accurate speech processing, selecting the most appropriate algorithms for feature extraction, and describing how these algorithms are optimised for real-time operation on a normal multimedia PC. Modules for speech level, pitch, vowel quality tracking, fricative detection and speech segmentation are presented. Tools and methods for implementing the graphics are also discussed.

Chapter 7 describes the results of the user trials conducted for evaluating the suitability and efficacy of the prototype system designed and implemented in Chapter 6. The assessment was carried out with the help of hearing-impaired speakers and therapists, at several locations in Britain. The subjects were hearing-impaired people belonging to the following categories: pre-lingually disabled, post-lingually disabled, elderly disabled, and those with cochlear implants. This assessment gave very good results on the effectiveness of the system, and produced a set of suggestions for future improvement.

CHAPTER 2

Characteristics of Hearing-Impaired Speech

2.1 Introduction

It is a sad fact that, in humans, a hearing problem is invariably accompanied by some problem with speech production. For this reason the speech of a hearing-impaired person can exhibit a variety of pathological problems which will normally result in a reduction in intelligibility. In order to understand how hearing impairment can affect the speech production system, the concept of vocal feedback, taken from information theory, can be used to model the way in which the output from a process or system can be used to dynamically influence the process or system itself. Take as an example a thermostat in a room heating system. The thermostat continuously measures the temperature of the air in the room, and when the temperature goes above a threshold, the heating system is turned off, and turned on again when the air temperature falls below another pre-arranged threshold. The output of the system, the heated air in the room, is continuously measured and information is fed back into the system in order to influence the behaviour of the system. Another example, where the system is a human, is the action of driving a car. The position of the steering

wheel is continuously corrected according to the perceived direction of motion of the car. In the case of speech production, vocal feedback works through the ears, facial bones and tissues, and is normally a continuous process.

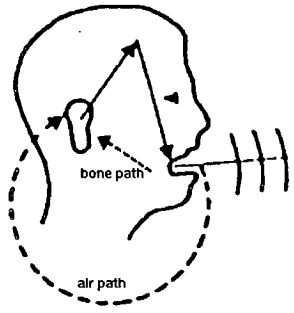


Figure 2.1. Vocal feedback

The process of vocal feedback which allows a talker to hear their own speech is part of the *speech chain*, the intrinsic connection between speech production and hearing (Denes & Pinson, 1963).

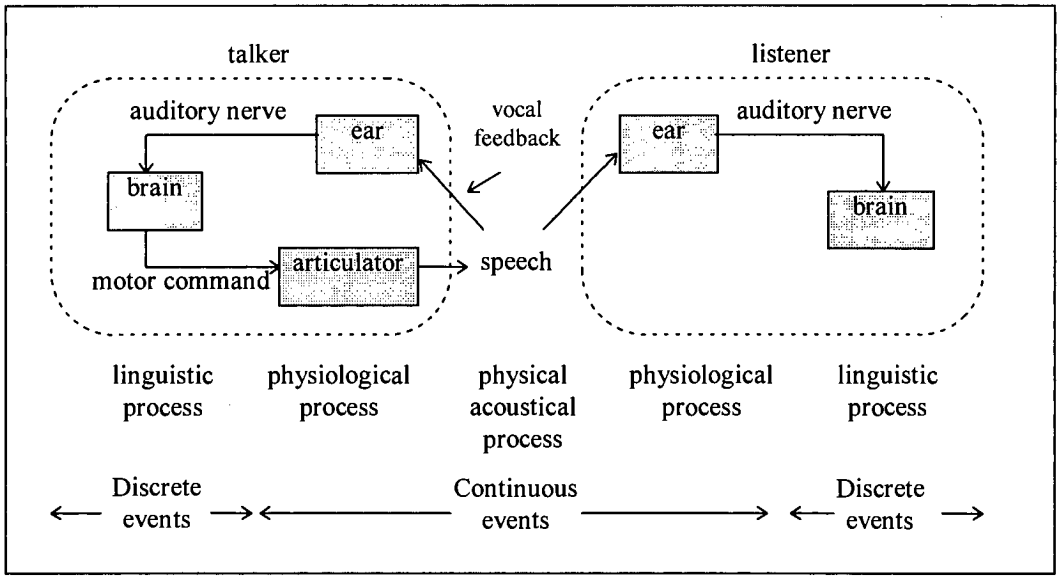


Figure 2.2. The speech chain

The process of vocal feedback is only partly attributable to conscious functions. An example of conscious feedback is when a talker corrects a defect in pronunciation, or a stumble in speaking, when hearing something which is at variance with the intended utterance. At a sub-conscious level, the process by which a talker hears their voice happens continuously. For example, speaking is difficult if the presence of loud noise prevents the talker from listening to their voice. Also, it can be difficult to speak during a trans-continental phone call, because of the delayed echo which is heard by the talker

if it is not properly suppressed. If the same effect is repeated in laboratory experiments and the delay is increased to a few tenths of a second, speaking is almost impossible. This demonstrates the important role played by vocal feedback, and how minor changes in the process cause very noticeable changes in speech production.

Auditory feedback is the most important process by which a talker monitors their own voice, but there are other modes that contribute to this process. One of these is the "hearing" of the voice through vibrations of the bones and the tissues of the skull. Actually the larger part of auditory feedback is via this means, and only partially via the air pressure wave that reaches the ears. This is evidences by the surprise shown by a speaker hearing the sound of their own voice from recordings since the direct feedback path is absent. A third mode is the kinaesthetic feedback that allows a talker to sense the movements and the position of the articulators tongue, jaws etc., and to sub-consciously use this information in maintaining the control of these articulators. When this information is disturbed, for example after an injection of anaesthetic for dental treatment or after a stroke, speech becomes slurred.

People who cannot hear their own voice because of hearing impairment (transmission through the bones of the skull needs the integrity of most of the ear), can only rely on kinaesthetic feedback. Such a reliance is inadequate for normal speech production but moreover when it is dissociated from the other modes of feedback, may lead to undesirable effects, as in Wirz (1987). As a result, intelligibility of the speech of hearing-impaired people is reduced (Hudgin & Numbers, 1942; Markides, 1970; Levitt & Nye, 1971) with the consequence that, among other things, a large number of people are prevented from participating in the social interactions of normal-hearing people (Lippmann, 1985).

This chapter deals with the process of speech production for both normal talkers and hearing-impaired talkers. Its purpose is first to analyse the voice production process in normal talkers, and then show how hearing impairment modifies this process. A wide range of theories which describe the speech production process are available. In addition to the classical Acoustical Theory of Speech Production (Fant, 1960), other theories on speech production are the Progressing Wave Model (Rabiner & Shafer, 1978), the Vocal Cord Model (Stevens, 1977), and the Articulatory Model (Shirai & Honda, 1980). For a comprehensive coverage of the normal aspects of speech, including physiological elements of speech production, and speech acoustics, see for example Minifie et al. (1973). The Acoustic Theory of Speech Production is judged most appropriate here because it gives the reader the appropriate background information for understanding the speech analysis algorithms used in this thesis - especially with reference to the content of Chapter 6. This theoretical development is then expanded to include a description of hearing impairments, clarifying how and why hearing impairments occur. Finally a description of the effects of deleterious hearing impairments on speech production highlights those aspects of hearing-impaired speech which are amenable to speech therapy methods.

2.2 Normal speech production

The normal speech production process can be thought as a chain of events originating in the brain, involving nerve impulses, muscular events and articulator movements, interactions between air pressure and air flow, and finally ending in the formation of an acoustic signal. The acoustic signal contains the relevant information on the linguistic content of an utterance plus information on the owner of the voice. This description of normal speech production focuses on the acoustic signal for two reasons: (1) the acoustic signal can be picked up with a single microphone, a convenient and non-invasive device; and (2) a theory exists as the Acoustic Theory of Speech Production (Fant, 1960), which describes, in terms of acoustic physics, the behaviour of the vocal organs and the sound produced from them. The description of normal speech production provided here will serve to provide the background knowledge which underpins the methods used for automatic analysis of speech features to be described in Chapter 6.

In this theory, the vocal tract is modelled as a system consisting of connected acoustic cavities excited by an acoustic source. Theories considering this aspect have been proposed since the eighteenth century, and early simulations were done by Krazenstein (1782) and Von Kempelen (1791). The modern theory is founded on the theoretical and experimental work by Fant (1960) with contributions from Flanagan (1965). The fast moving pace of results in this field have been enabled by developments in fields such as electronics, digital computers, radiography, fast radio-cinematography, and laryngoscopy. According to this theory, speech is the result of changes to a signal produced by the excitation source after traversing the acoustic filter.

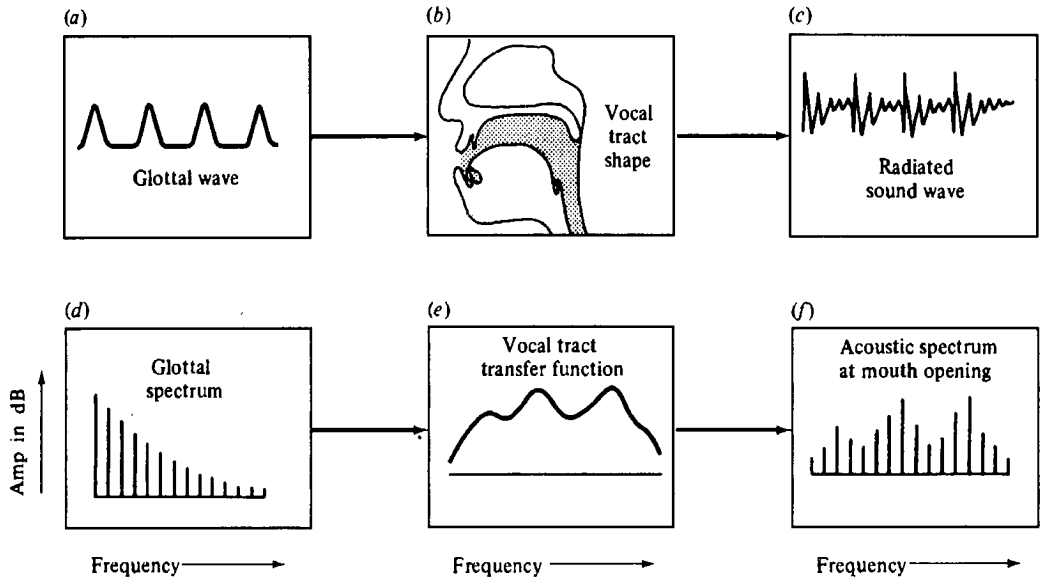


Figure 2.3. Physiological and acoustical characteristics of speech sound production

Figure 2.3 shows how voiced (vowel) sounds are produced (after Fant, 1960). The sound source (vocal chords) emits a quasi-periodic signal (glottal wave) which is modified by the shapes of the supra-glottal cavities and the positions of articulators. The dimensions of these cavities change according to the vowel being produced. In the case of consonants, a noise source is involved, which may substitute or co-exist with the glottal source. Some consonantal sounds, such as plosives (the /p/ in "papa") and affricates (the /f/ in "father") are produced by rapid changes in volume of the supra-glottal cavities. In both cases, the sound produced depends on the source characteristics (voiced or unvoiced) and on the changes due to different acoustic configurations of phono-articulatory cavities, and also on the rate at which these changes occur.

2.2.1 Vowel production

The source

When the respiratory muscles associated with the lungs are caused to contract, the air pressure in the lungs increases and provokes a high pressure air stream through the cavities of the vocal tract. If the muscles surrounding the vocal chords cause them to vibrate, then the pressure of the superglottal air stream will vary periodically, in synchrony with the opening and closing of the vocal folds. In the case of whispered vowels, the vocal folds stay in a position in which the glottal orifice is greatly reduced in size, functioning in this way as a form of noise source. Studies on the causes of the vibration of the vocal folds have given contradictory theories. The "mio-elastic" theory (which is currently accepted) derives from the work of Smith (1954), J.W. Van Den Berg (1954), F. Faaborg-Andersen (1957), and B. Sonesson (1960). According to this theory, the vibrations of the vocal folds depend on mechanical causes that are a function of both the subglottal air pressure and the degree of tension of the vocal folds. The larynx operates as a (relaxation) oscillator, consisting of a source of energy (the lungs) together with an oscillating device (the vocal folds) which have no natural oscillation frequency. The combined effect of these two elements determines the oscillation frequency of the vocal folds.

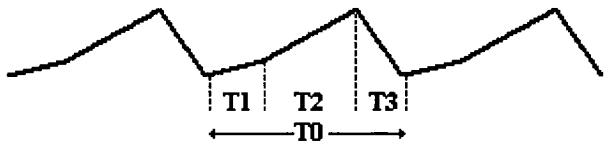


Figure 2.4. Glottal airflow

The airflow through the vocal folds varies with a waveform having the shape of a 'sawtooth' and during a period T0 (see Figure 2.4) it is possible to identify three phases, abduction (whose duration is T1), opening (T2) and adduction (T3).

The mechanical explanation of the operation of the vocal folds involves the cyclic variation in force that the air in the lungs exerts on the glottis. As soon as the air pressure opens the vocal folds, an acceleration of the air stream takes place in the glottis, and this causes a resultant decrease in the air pressure (a phenomenon called "The Bernoulli effect"[1]), and the vocal folds tend, because of their own muscular tension, to come closer until they snap into complete closure, at which point the system returns to the starting point of another cycle. When the talker seeks to talk more loudly, the sub-glottal air pressure increases, the voice intensity increases and the oscillation period T_0 decreases, unless compensated by a weakening of the muscular tension of the vocal folds. The result is an increase of the fundamental frequency (F_0) of the speech utterance. The vocalised source (when the vocal folds oscillate) exhibits a line spectrum in which the harmonic components are multiples of the fundamental frequency. The spectral envelope shows an average attenuation of 12 dB/octave (see Figure 2.5).

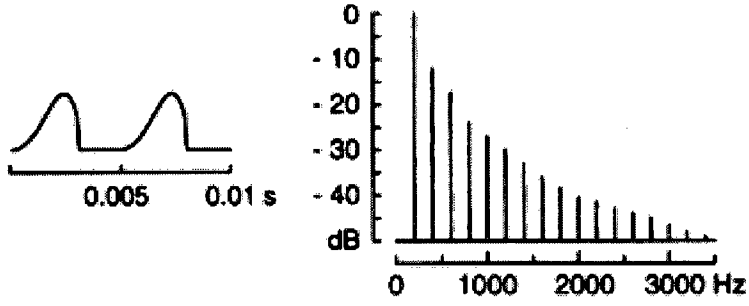


Figure 2.5. Spectral characteristics of the glottal wave

[1] First derived (1738) by the Swiss mathematician Daniel Bernoulli, the theorem states that the total mechanical energy of a flowing fluid, comprising the energy associated with fluid pressure, the gravitational potential energy of elevation, and the kinetic energy of fluid motion, remains constant. Bernoulli's theorem is the principle of energy conservation for ideal fluids in steady, or streamline, flow. Bernoulli's theorem implies, therefore, that if the fluid flows horizontally so that no change in gravitational potential energy occurs, then a decrease in fluid pressure is associated with an increase in fluid velocity. If the fluid is flowing through a horizontal pipe of varying cross-sectional area, for example, the fluid speeds up in constricted areas so that the pressure the fluid exerts is least where the cross section is smallest. This phenomenon is sometimes called the Venturi effect, after the Italian scientist G.B. Venturi (1746-1822), who first noted the effects of constricted channels on fluid flow. (*From Encyclopaedia Britannica)*

The vocal tract when producing vowels

The vocal tract during the production of a vowel can be described in terms of an Helmoltz resonator[1] or two resonators coupled together, and excited by a sound source S_1 (see Figure 2.6).

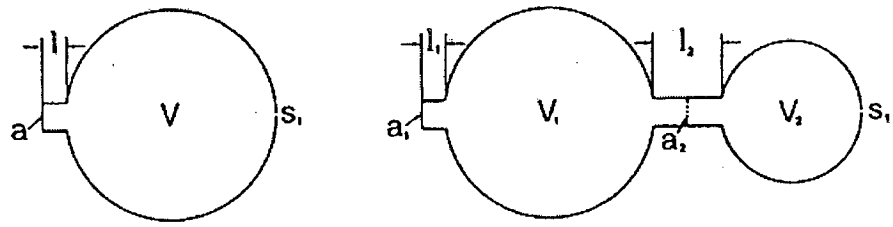


Figure 2.6. Helmholtz resonator: single and coupled

This simple model is only valid for harmonics of the laryngeal signal below 300 Hz, since the supra-glottal cavities tend to introduce resonances at higher frequencies. The model approximation with two resonators coupled together, though better, also shows limitations that cannot be neglected. The modern acoustic theory of speech production approximates the configuration of the vocal tract to an acoustic duct with radial symmetry, whose transverse section has a variable area. To obtain such a representation, the typical procedure is to start from a X-ray picture of the vocal tract. A line through the centre of the cross sectional area of the vocal tract is drawn (see Figure 2.7a). Then the shape of the cross sections are determined at various positions of the vocal tract, using X-ray pictures taken perpendicularly to each position, or any other available data (pictures from the front, endoscopy, etc. (see Figure 2.7b.). After thus finding the areas of the transverse sections, the continuous area function over the full length of the vocal tract can be interpolated (Figure 2.7c).

[1] A Helmholtz resonator consists of a hollow metallic sphere, having a volume of V, with a hole that links the sound source with the inside of the cavity, and with a nozzle, having a section of a and length l, and allowing the sound to be radiated outside. Excited by the source S_1, the resonator is used to produce sounds. The theory shows that in a resonator, the intensity of certain harmonic components of the signal is reinforced and, for a given section of the opening, the value of their frequencies is an inverse function of the square root of the resonator's volume. This means that the resonance frequency is independent of the shape of the cavity provided that the cavity maintains the same volume.

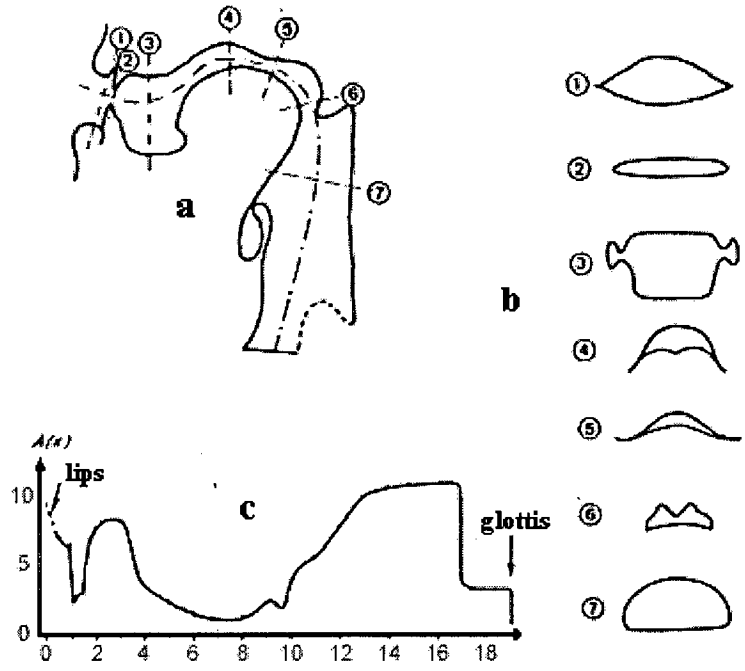


Figure 2.7. Median saggittal section, transverse sections, area function (from Fant, 1960)

For simulation purposes, the continuous shape of the area function can be quantized in short segments with constant length. The vocal tract model after this procedure is equivalent to a non-uniform acoustic duct consisting of circular sections along a straight axial line, and as such it can be studied using techniques of electro-acoustic analysis.

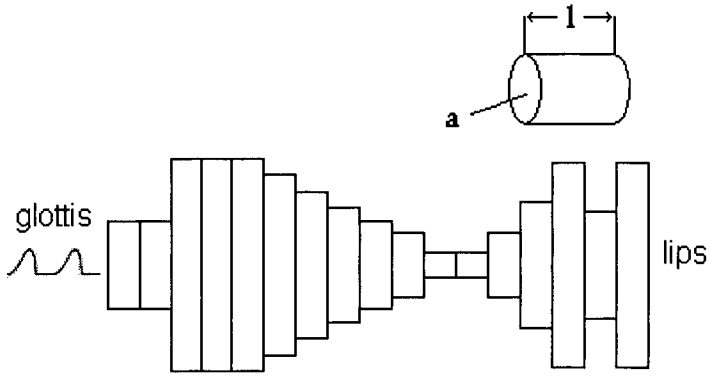


Figure 2.8. Acoustic section model of the vocal tract

In particular, the vocal tract is considered as a succession of elementary pipes, each having an appropriate length l and cross-section a, travelled by longitudinal waves (see Figure 2.8).

Using analogies between acoustical and electrical parameters for each elementary pipe in which an acoustic wave is propagating, it is possible to construct an equivalent electric circuit where a corresponding electric current flows. From the series of elementary electrical circuits (representing the pipes that schematise the vocal tract), an electrical analogue of the vocal tract is obtained, and from which it is possible to determine the resonant characteristics of the tract itself. In fact, by exciting the electrical analogue of a vocal tract with a sinusoidal signal having constant amplitude, and variable frequency, it is possible to measure the intensity of the output signal at each value of the input frequency, and thereby possible to draw the frequency response (or transfer response) of the vocal tract.

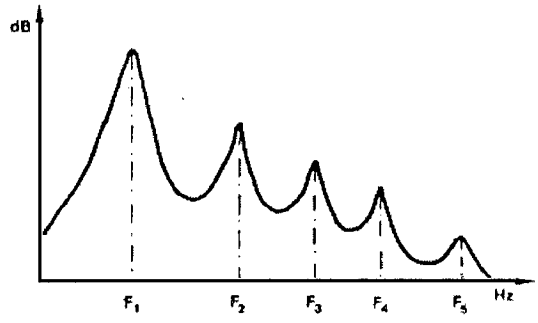


Figure 2.9. Example of frequency response of the vocal tract

As can be seen from Figure 2.9, the vocal tract frequency response exhibits a series of peaks in intensity corresponding to the frequency values indicated by F_1, F_2, F_3, F_4... that constitute the resonance frequencies of the whole vocal tract. The vocal tract therefore has the property of enhancing the intensity of some frequencies and reducing others, working as a complex resonator. As stated above, the greater the number of elementary pipes (and therefore of analogous electric circuits), the better the acoustic approximation. However the use of a large number of elementary circuits involves a increased computational complexity so the issue is to find a good compromise between quality of approximation and complexity. A model based on two or four pipes can deliver satisfactory results for most non-nasal vowels. The neutral vowel (schwa) can be modelled with only a single pipe having a constant section. Using a four pipe model, Fant (1960) has built "monograms" to record the first five resonances (formants) as a function of the real dimensions of the vocal tract, drawn from X-ray photographs.

To summarise, the transfer function of vocalised vowels shows a series of peaks corresponding to certain frequencies named resonance frequencies, and labelled as F_1, F_2, F_3... These resonances are not the resonance frequencies of specific cavities, but the result of the whole set of cavities that constitute the vocal tract.

The vocal tract when producing nasal vowels

The lowering of the velum causes both the vocal and nasal cavities to create resonances, and radiate acoustic energy simultaneously from lips and nostrils with most of the sound energy radiated from the lips. Inclusion of the nasal cavity modifies the frequency response of the corresponding non-nasal vowel, causing a shift in the resonance frequencies and the appearance of new resonances and anti-resonances. Although nasalization has been the subject of many studies by phoneticians and experts in speech analysis-synthesis and speech recognition (see for example Rooney, 1990), it is not an easy phenomenon to study, and it may cause errors in the interpretation of results from such electrical models.

2.2.2 General considerations on the acoustic characteristics of vowels

Frequency

Knowing the characteristics of the excitation source, the transfer function of the vocal tract, and the effects of lip radiation, it is possible to deduce the characteristics of the sound produced. As the spectral envelope of the glottal signal is uniformly decreasing at 12 dB/octave, and the effect of the radiation from the lips corresponds to a spectral rise of 6 dB/octave, the resonance frequencies of the vocal tract appear superimposed with a resulting -6 dB/octave slope in the speech signal. Harmonics of the excitation source having frequency near F_1, F_2, F_3... become amplified, while other frequencies become more attenuated since they are far from these frequencies or are in anti-resonances. These peaks of intensity in the spectral envelope are named formants and labelled as F_1, F_2, F_3, etc.; each formant is characterised by its frequency F_n, its bandwidth B_n and by its amplitude level L_n (n is the formant number). The spectral envelope is more defined as the number of lines in the spectrum increases, and the laryngeal frequency of the speaker correspondingly decreases. For this reason it is easier to determine the value of the formant frequencies for male speech than for female speech. It should be noted that the formants are not harmonically related to the source excitation frequency. Formant frequencies are a function of the vocal tract dimensions. For this reason there are differences for the formant values of the same vowel produced by a man, a woman and a child. The increase in formant frequency between man and woman for F_1 vary between 5% and 30%, for F_2 between 10% and 25%, and for F_3 between 10% and 20%; those between adults and children are even bigger (Peterson & Barney, 1952).

Pickett (1980), combining results from Dunn, Fant and others, has summarised some practical rules which relate formants and vocal tract:

- Length rule: The average formant frequencies for a vowel are inversely proportional to the length of the vocal tract (this accounts for the up-shift of formants from men to women to children).
- Lip-rounding rule: The frequencies of all formants are lowered by lip-rounding; the higher the lip rounding, the lower the formant frequency.
- Area Ratio rule: Formant frequencies are almost independent of the actual cross section. Formant frequencies can be scaled up or down, provided that the ratio of formant frequency remain unchanged.
- Constriction rule: The first two formants are shifted, relatively to their neutral position, because of the narrow constriction in the oral cavity due to the tongue. Note that the tongue position does not influence F_3.

The first two or three formant frequencies give an acoustic description which is adequate for characterising the phonetic quality of a vowel. Often phoneticians use only the frequencies of the first two formants, which are enough to characterise almost all vowels.

Amplitude Level

The intensity of the harmonics of the glottal source shows an average attenuation of 12 dB/octave (see Figure 2.5), so the intensity of higher frequency formants tends to be lower. This attenuation is partially compensated by an increase of 6 dB/octave due to the effect of the lip radiation. The change in absolute intensity of the first formant is relatively small among different vowels (about 4 dB). For this reason phoneticians often calculate the relative intensity of a formant with reference to the intensity of the first formant. In this way the differences L_2-L_1, L_3-L_1, and L_4-L_1 are obtained. The difference L_2-L_1 varies between -5 dB for /ɔ/ (as in "bought") and -28 dB for /i/ (as in "beat"); L_3-L_1 varies between -18 dB for /ɛ/ (as in "bet") and -40 dB for /u/ (as in "boot"). The intensity of the third formant in back vowels (/ɑ/ as in "father", /ω/ as in "good", /u/ as in "food") is very weak, so that it is often difficult to give it prominence in analysis. To determine the intensity of a formant both the frequency value and also the -3dB bandwidth B_n of the formant must be measured. Fundamental studies on formant bandwidth were made by Fant (1956) and Dunn (1961). Formant bandwidth normally varies between 40 Hz and 240 Hz: it increases with formant frequency. Among individual talkers, the range of variations in formant intensity is relatively wider than the changes in formant frequency.

To summarise, formant intensity is weaker for higher formant frequencies (with the exception of /i/ as in "beat"). Formant intensity is stronger when formants are closer together. In fact, if F_2 is close to F_1, the low character of the voice is reinforced, because L_1 and L_2 are reinforced, and L_3 attenuated. In

contrast, if F_2 is close to F_3, the sharp character of the vowel is reinforced, because L_1 is attenuated and L_2 and L_3 reinforced.

Nasal Vowels

The acoustic characteristics of nasal vowels have been studied by Smith (1951), Delattre (1954, 1965), House-Stevens (1956), Kurowski & Blumstein (1984), Rooney (1990), and Harrington (1994). Comparing the spectral envelope of a nasal vowel with that for a corresponding non-nasal vowel, shows:

- no significant change in the fundamental frequency
- appearance of a first nasal formant FN_1 having a frequency of around 250 Hz
- tendency for F_1 to average 500 Hz
- strong attenuation of L_1
- appearance of a second nasal formant FN_2 around 1000 Hz
- no significant change of F_2
- appearance of a third nasal formant FN_3 around 2000 Hz
- slight shift of F_3 and F_4

Also, the time duration of nasal vowels is generally longer than for the corresponding non-nasal vowels. While all researchers agree on the fact that nasal coupling widens resonances and introduces spectral zeros, there is less agreement on the frequencies of these zeros. Flanagan (1972) shows them around 1300 Hz, while Pickett (1980) at 600 Hz and 2 kHz. Fujimura (1962) show them as appearing at a wide range of frequencies.

2.2.3 Consonant production

Plosives

Voiced plosives are characterised by two distinct phases: closure and release. The first phase, where the articulators meet to form a complete or near complete closure in the vocal tract, varies in length according to the sound produced but is generally short. Voicing may persist throughout the closure, but the longer the closure the more likely that voicing will not be maintained, since airflow across the vocal folds is necessary to support their vibration. Air pressure builds up behind the point of closure in the vocal tract and when it is released, there is a sudden outflow of air, called the release burst. The airflow quickly reduces and the turbulent, high-energy burst gives way to regular voicing.

Similarly, voiceless plosives have a closure and release phase, although with these sounds, there is no voicing during the closure phase and voicing only reappears some time after the release burst. The duration of this voice onset delay (also known as the aspiration phase) varies within and across languages. For example, in English the plosives /p/, /t/, /k/ have much longer aspiration phases than they do in French or Italian. However, when following fricatives in a consonant cluster, these sounds have almost no voice onset delay in English ("port" vs. "sport", "tar" vs. "star", "key" vs. "ski").

Fricatives

Theoretical studies of the production and acoustic qualities of fricatives have proven to be complex. In the case of voiced fricatives such as "v" and "z", a periodic sound source from the vibrating vocal folds excites the vocal tract, and then a noise source is added at the point of constriction in the vocal tract, where the airflow is turbulent. Spectrally, voiced fricatives are characterised by both a periodic and a noise component, and generally low amplitude. Calculation of the electrical analogue is not easy, because it is difficult to measure accurately the dimensions of the constriction, and the spectral characteristics of the noise source are not thoroughly known. Voiceless fricatives such as "f", "sh" and "th" are characterised only by turbulence in the airstream and the point of narrowest constriction, causing audible frication. They generally have higher amplitude than their voiced equivalents. Fricatives may be changed in duration, over a reasonable range, without affecting their phonemic identification. The following Table 2.1, providing physiological descriptions and information on source spectrum and output spectrum, was extracted from Fant (1960), Strevens (1960) and Heinz & Stevens (1961).

Physiological description	*Source spectrum*	*Output spectrum*
Interdental		
/θ/ (as in "thin") and /ð/ (as in "that") are produced with the tip of the tongue close to, or touching, the inner edge of the upper incisors. Air is forced through a narrow slit between the bottom of the upper teeth and the top surface of the tongue. The broad width to height nature of the orifice shape yields a low intensity level, broad bandwidth of noise.	Flat noise spectrum from about 1000 Hz to 10000 Hz. Energy may drop off at about 3 dB/octave.	Strevens observes that /θ/ has low intensity noise with the highest center of gravity in the frequency domain of any English fricative. Largest amplitudes of energy are from 7000 to 8000 Hz. Heinz and Stevens found that listeners identified wide band resonances from 6500 to 8000 Hz as either /f/ (as in "farm") or /θ/ (as in "thin").

Labiodental		
/f/ (as in "farm") and /v/ (as in "very") are produced with the upper teeth close to the inner surface of lower lip. The air stream passes between the teeth and the lower lip, and also through some of the interstices between the upper teeth. The broad width and low height to the elliptical orifice offers a large resistance to the outward flow of air, as well as causing a low intensity, wide band noise to be produced. The noise source is relatively unmodified by the vocal tract since the noise source is located near the output of the resonance tube.	A low intensity noise band from 800 to 10000 Hz. Amplitude of noise drops approximately 3 dB to 6 dB / octave	Low intensity noise ranging from approximately 1500 Hz to 7500- 8000 Hz. Strevens has identified low level resonances at 1900, 4000 and 5000 Hz. Fant suggests the major resonance will occur at 6000 - 7000 Hz and is dependent on the resonance of the air column in the constriction and shallow cavity formed by the lips in front of the upper incisors. Fant suggests that this high frequency resonance will be very low in amplitude and that the /f/ sound is perhaps better described by a broad band noise with no observable resonances.
Lingua-Alveolar		
/s/ (as in "kiss") and /z/ (as in "cousin") are produced with the tongue tip or tongue blade raised to approximate the alveolar ridge. The tongue is grooved forming a narrow air channel down the center of the tongue. The closer approximation to a circular orifice provides a more efficient conversion of aerodynamic power to acoustic power than does the wide but low elliptical orifice required for /f/ (as in "farm") or /θ/ (as in "thin"). Turbulence is generated at the constriction and also at the cutting edge of the upper incisor teeth.	The sound source spectrum is flat from 300-4000 Hz followed by a 6 dB/octave drop above 4000 Hz.	Due to an antiresonance around 3500 Hz, very little energy is observed in the output spectrum below 4000 Hz. No characteristic resonance pattern but usually a major energy peak between 4000 and 7000 Hz.
Lingua-Palatal		
/ʃ/ (as in "sugar") and /ʒ/ (as in "beige") are produced with the blade of the tongue, or the tip and blade approaching the palate approximately where the alveolar ridge joins the hard palate. They require the tongue to be only slightly grooved, providing a larger area for turbulence. The air stream is set into turbulence in the constriction and possibly at the teeth.	Approximately a flat source spectrum (0 dB/octave) from 300-6000 Hz.	Lowest output energies around 1600-2500 Hz. Very sharp cut-off of high frequency energies around 7000 Hz. Most of the fricative energy is in the lower frequencies. Strevens (1960) and Heinz and Stevens (1961) show the first two resonances occurring at about 2500 and 5000 Hz respectively.

Glottal		
/h/ (as in "hat") is produced by increasing the airflow through the larynx and creating turbulence within a partially constricted glottis.	Broad spectrum noise.	The output spectrum ranges from about 400 Hz to 6500 Hz. Several peaks occur in the output spectrum for /h/, one around 1000 Hz and another around 1700 Hz. Since the whole vocal tract resonates during /h/, lower frequency energies are resonated than for fricatives with more forward places of articulation. There are no physiological constraints placed on the tongue during /h/, providing maximal coarticulation with the following vowel. Hence, the resonances for /h/ may closely approximate those for the following vowel.

Table 2.1. Acoustical properties of English fricatives (from Minifie et al., 1973)

Affricates

Affricates, for example /t͡ʃ/ as in "church", /d͡ʒ/ as in "judge", are similar to plosives in that they have a closure and a release phase. However, with these sounds, the release is near the place of articulation of the closure, causing audible frication.

Approximants

With approximants, (/l/, /r/, /w/, /j/) as with vowels, a voiced source excites the vocal tract. Approximants are characterised by faster transitions than vowels. In the articulation of /l/ (as in "ball") there are two lateral passages of air, symmetrical in respect of the top of the tongue. These passages connect the internal cavities with the cavities in front of the tongue; note that in some talkers /l/ is unilateral). This complicates the study of such a configuration, even if the longitudinal dimension retains primary importance. The spectrum shows lines that have a formant structure. There are numerous configurations of /r/ (as in "rat"), relative to the place of articulation, which may involve the tip of the tongue and the alveolar area (the area of the palate just behind the upper teeth), or the back of the tongue and the uvular and velar area (areas at the back of the mouth).

Nasal Consonants

Nasal consonants are voiced, so the glottal source excites the three cavities: pharyngeal, vocal and nasal. During the closure of the mouth, the sound signal is only radiated by the nostrils.

The frequency response obtained by an electrical analogue of the set of cavities shows the following characteristics:

- a series of resonance frequencies FN_1, FN_2, FN_3, etc. corresponding to the intensity peaks LN_1, LN_2, LN_3, etc. of the nasal formants. FN_1 is rather low, since the volume of the set of cavities as a whole is greater than for non-nasal vowels
- appearance of anti-resonances, as distinct minima in the frequency response. These alter the intensity ratios among formants (but LN_1 retains greater intensity). Theoretically, the presence of these anti-resonances is explained by the fact that the nasal cavity is in parallel with the oral cavity
- weakening of the global intensity of the consonant, due to the fibrous nature of the walls of the nasal cavity

To summarise, the acoustic characteristics of a nasal consonant are:

- a spectrum showing lines, since it is a vocalised emission
- a formant structure which is similar for the sounds /m/ (as in "comb"), /n/ (as in "ran") and /ŋ/ (as in "bank") showing:

  FN_1 at around 250 Hz

  FN_2 between 1000 and 1200 Hz

  FN_3 between 1800 and 2300 Hz
- a global intensity which is weaker than its non-nasal equivalent.

The sudden opening of the vocal cavity causes a quick shift of the anti-resonances, with a consequent change in the spectrum, and a simultaneous raising of the intensity of the formants that depend on the vocal tract. These are the most important changes for perceptual discrimination among the different nasal consonants.

2.2.4 Normal values

As a reference, graphics showing normal values for vocal intensity, fundamental frequency and vowel quality are presented in the following sections.

Intensity

The curve in Figure 2.10 represents the amplitude density distribution function, calculated for utterances from 80 speakers (4 speakers x 20 languages) having a duration of about 37 minutes (Irii et al., 1987). The horizontal axis is normalised by the long-term effective value. The curve shows that the dynamic range of speech amplitude exceeds 50 dB. Results from the same study show that the mean value for male voices is about 4.5 dB higher than that for female voices. Also, the long-term effective value under an high-noise level condition is raised according to that noise level. Age has some effect on the maximum sound pressure level, with young adults exhibiting a level which is on the order of 6 dB (i.e. double) higher than elderly people.

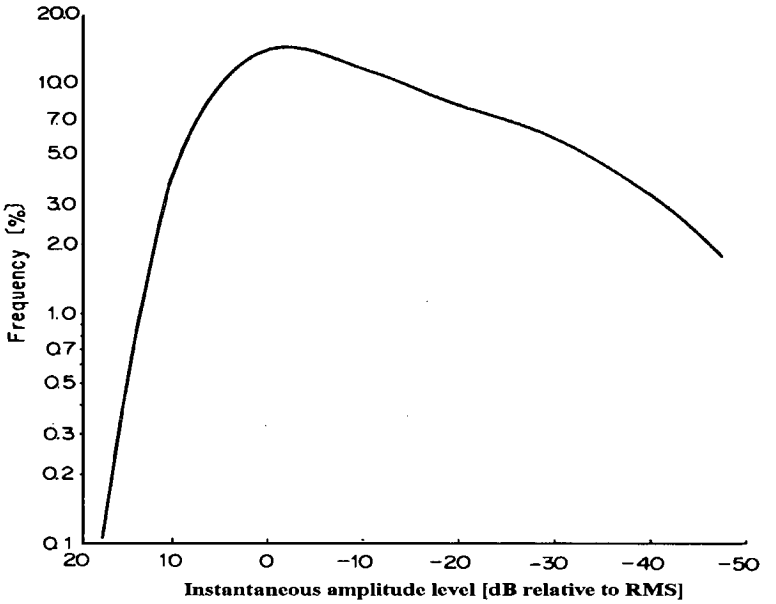


Figure 2.10. Amplitude density distribution (from Irii et al., 1987)

In another study, Coleman et al. (1977) assessed the sound pressure level profiles of 10 men, age 21-34 years, and 12 women, age 20-39 years. The subjects produced sustained vowels for at least 2 seconds, at 10% intervals of their pitch range. The curves in Figure 2.11a (men) and 2.11b (women) show the range of sound pressure level as a function of fundamental frequency level. The lower curves (minimal sustainable intensity) are produced by instructing the subjects to say the vowel as

quietly as possible without whispering; the upper curves (maximal intensity) are generated by asking the subjects to shout the vowels as loud as possible but without screaming or squealing.

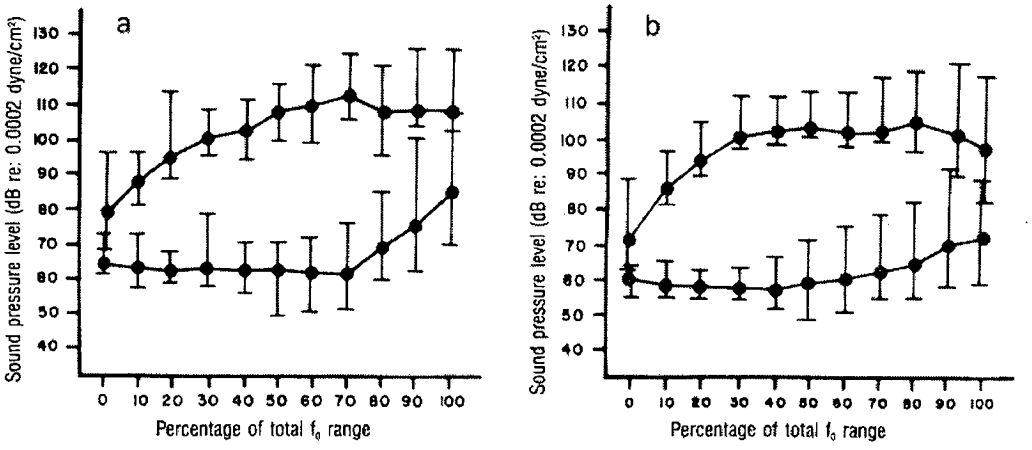


Figure 2.11. Range of sound pressure level (average, min, max) as a function of fundamental frequency level for (a) men and (b) women (from Coleman et al., 1977)

Figure 2.11 suggests that the sound pressure level tends to increase as the fundamental frequency (F_0) rises.

Fundamental frequency

An extensive description of normal and pathologic vocal fundamental frequency may be found for example in Baken (1987). The following brief description on variations in fundamental frequency is an extract from Furui (1989), and is more appropriate for the purpose of this chapter: "Statistical analysis of temporal variation in fundamental frequency during conversational speech for every speaker indicates that the mean and standard deviation for female voices are roughly twice those for male voices. The fundamental frequency distributed over speakers on a logarithmic frequency scale can be approximated by two normal distribution functions which correspond to male and female voices, respectively (see Figure 2.12). The mean and standard deviation for male voices are 125 and 20.5 Hz, respectively, whereas those for female voices are two times larger. Intraspeaker variation is roughly 20% smaller than interspeaker variation."

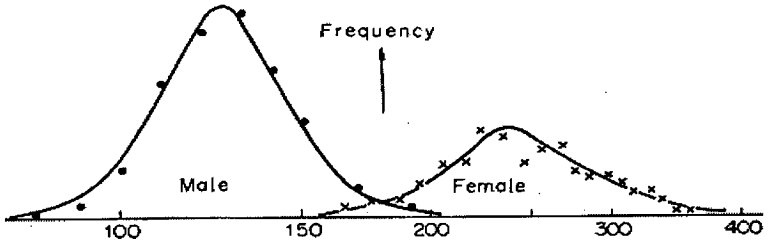


Figure 2.12. Fundamental frequency distribution over speakers (from Furui, 1989)

Vowels

Figure 2.13 shows the frequency distribution of the averaged first and second formants for English vowels produced by men, women and children [Figure taken from Handbook of Speech Pathology and Audiology, edited by Lee Edward Travis, 1971]

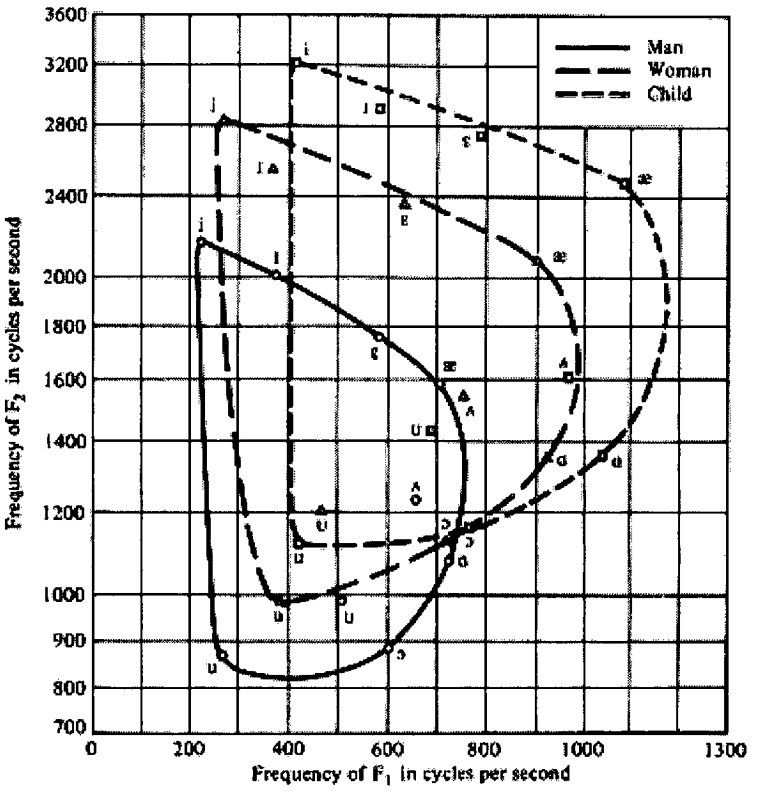


Figure 2.13. Distributions of the averaged first and second formants

Figure 2.14 shows the scatter diagram of formant frequencies of 10 English vowels produced by 76 speakers (33 adult males, 28 adult females and 15 children) measured by Peterson and Barney (1952), with examples of words using the vowels.

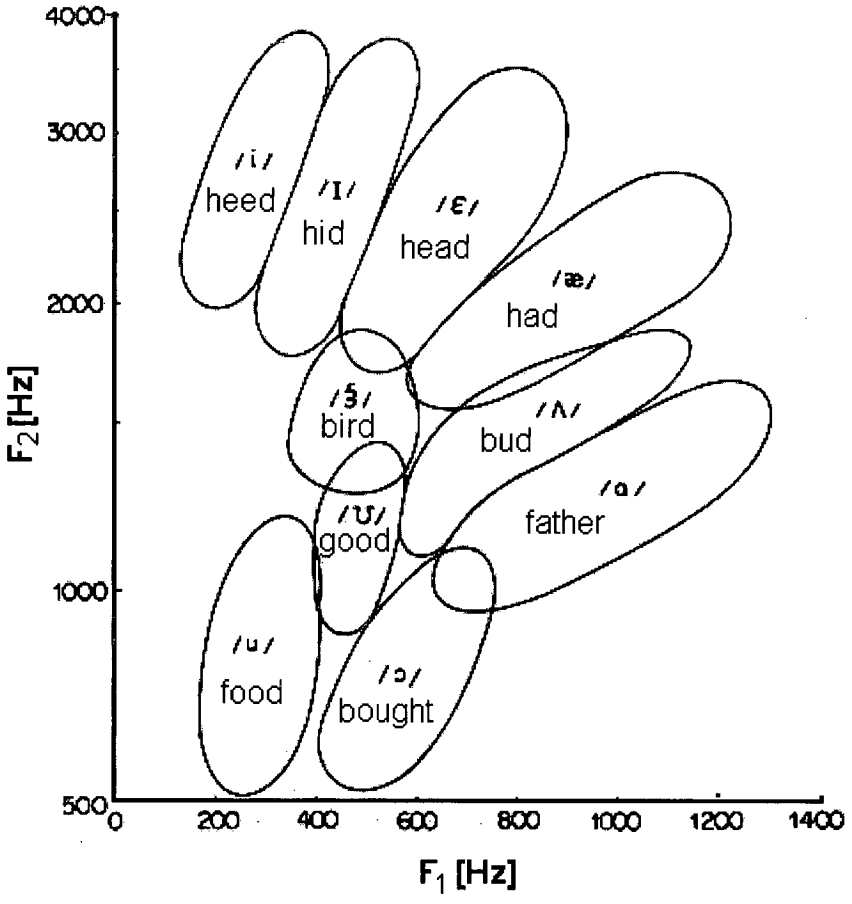


Figure 2.14. Scatter diagram of 10 English vowels, with examples (from Peterson & Barney, 1952)

Figure 2.15. shows the tongue and lip position for the following English vowels:

(1) [i] heed; (2) [ɪ] hid; (3) [ɛ] head; (4) [æ] had; (5) [ɑ] father; (6) [ɷ]good; (7) [u] food.

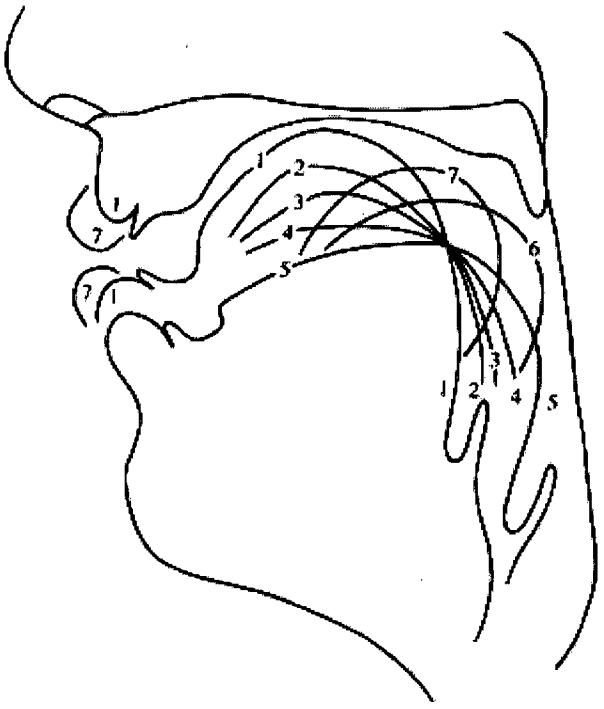


Figure 2.15. Tongue and lip position for certain English vowels: 1 heed, 2 hid, 3 head, 4 had, 5 father, 6 good, 7 food. The lip position for vowels 2,3 and 4 are in between those shown for 1 and 5. The lip position for vowel 6 is between those shown for 1 and 7 (from Ladefoged, 1982).

2.3 Characteristics of Hearing Impairments

2.3.1 Figures for Great Britain and the rest of Europe

Hearing impairment is a condition which is often under-estimated. It is also difficult to derive exact figures about the number of hearing-impaired people, since many people do not register their impairment. The figures detailed below can only be estimates about the overall picture in Great Britain and the rest of Europe[1].

Great Britain: Adults

Estimates of the number of adults in Great Britain in different hearing loss categories. These categories are defined as follows:

Mild Hearing Loss (25 - 40 dBHL) - have some difficulty in following speech mainly in noisy situations. Some wear hearing aids, others lip read.

Moderate Hearing Loss (41 - 70 dBHL) - find it difficult to follow speech without a hearing aid and have even greater difficulty in noisy situations. Most use a hearing aid or lip read and can use a telephone with an amplifier and/or inductive coupler and hearing aid.

Severe Hearing Loss (71 - 95 dBHL) - difficulty in following speech, even with a hearing aid many have to lip-read. Those deaf from birth often use sign language.

Profound Hearing Loss (96+ dBHL) - hearing aids of little benefit. Likely to use sign language.

Description of Hearing Loss	Number	Percentage of total adult population
Mild hearing loss	5.0 million	11.33%
Moderate hearing loss	2.2 million	4.99%
Severe hearing loss	0.24 million	0.54%
Profound hearing loss	0.06 million	0.14%
Total	7.5 million	17%

Table 2.2. Estimates of the number of Adults in Great Britain in different hearing loss categories

[1]Source for the figures on Great Britain: Royal National Institute for the Deaf, with assistance from the British Association of teachers of the Deaf, British Deaf Association, MRC Institute of Hearing Research, National Deaf Children's Society (1989). Source for the figures on Europe: Dr Adrian Davis, Medical Research Council, Nottingham. Audiology in Europe meeting, September 1992.

The following figure graphically represents the data listed above.

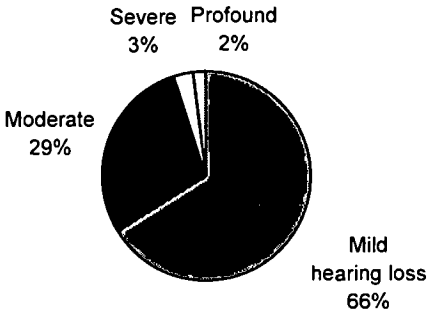


Figure 2.16. Hearing loss in hearing-impaired adults (Great Britain)

Of the 7.5 million hearing impaired:

- 5.625 million adults who suffer from hearing loss are over 60.
- Between the ages of 61 and 70 at least a third of people have some degree of hearing loss and at least 10% have a loss which is moderate or worse.
- In the 71-80 age group more than 50% of people have some degree of hearing loss and at least 20% have a loss which is moderate or worse.

The vast majority of people with hearing difficulties are elderly hard-of-hearing people. The following chart summarise the percentage of hearing loss according to the age.

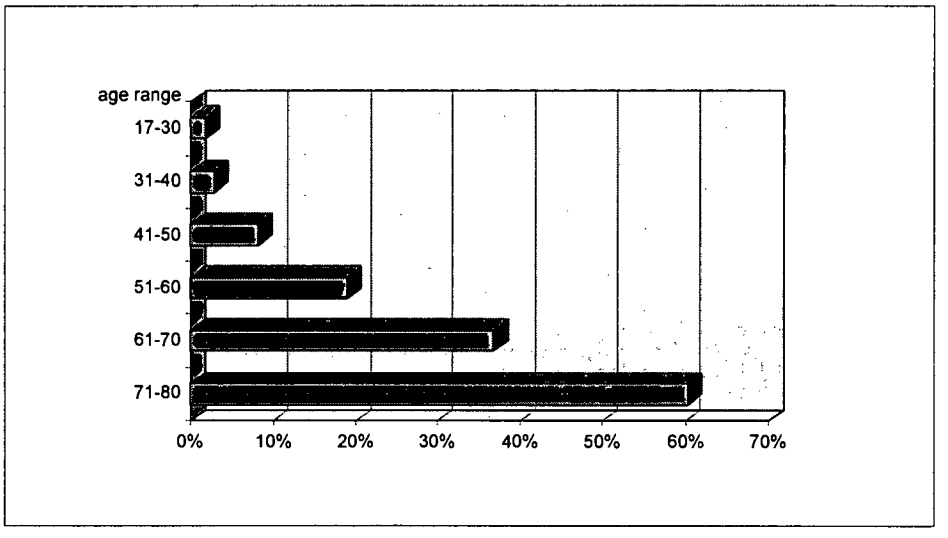


Figure 2.17. Percentage of hearing loss by age (Great Britain)

Great Britain: Children

Three in every thousand children born in the UK are born with some degree of hearing impairment, one in every thousand with severe or profound hearing loss. This means that approximately 100,000 children have some level of hearing impairment. By the age of 16 the prevalence of hearing difficulties has risen to six in every thousand. In addiction to this, many children suffer from ailments such as glue ear which most will grow out of but can affect speech in the early stages. It is estimated that there are approximately 25,000 profoundly pre-lingually deaf children in Great Britain.

Europe: Adults

The table below shows the number of adults, in areas of Europe, who have a certain degree of hearing loss.

Area	Population (million)	Mild (25+dBHL)	Mild-Moderate (35+dBHL)	Moderate-Worse (54+dBHL)
N. Europe	93.00	13.50 (14.5%)	7.58 (8.1%)	3.96 (4.2%)
W. Europe	178.00	25.02 (14.0 %)	13.81 (7.7%)	7.29 (4.0%)
E. Europe	96.00	11.39 (11.8%)	6.12 (6.3%)	3.17 (3.3%)
S. Europe	144.00	18.56 (12.8%)	10.01 (6.9%)	5.15 (3.5%)
Total	511.00	67.80 (13.2%)	37.09 (7.2%)	19.34 (3.7%)

Table 2. 3. Hearing loss in hearing-impaired adults (Europe)

Europe: Children

The table below shows the number of children throughout Europe that are aged over 5 with hearing impairment of 50 dBHL or worse in their better ear, the number of these children that are born profoundly deaf and the number that become profoundly deaf by the age of 5.

Area	Population (million)	All: 50 dBHL or worse by the age of 5 (thousands)	Of which: profoundly deaf at birth	Profoundly deaf by 5
N. Europe	93.00	1575.42 (1.69%)	351.54 (0.37%)	481.74 (0.51%)
W. Europe	178.00	2584.56 (1.45%)	576.20 (0.32%)	790.32 (0.44%)
E. Europe	96.00	1510.08 (1.57%)	336.96 (0.35%)	461.76 (0.48%)
S. Europe	144.00	1749.66 (1.21%)	390.94 (0.27%)	535.02 (0.37%)
Total	511.00	7419.72 (1.45%)	1655.64 (0.32%)	2268.84 (0.44%)

Table 2.4. Hearing loss in hearing-impaired children (Europe)

2.3.2 Causes of deafness

It is important to identify the causes and severity of the problem, especially with children where it has to be decided where and how they will be educated. The *otologist* is responsible for establishing where the problem is located. The *audiologist* is responsible in establishing the severity and type of hearing loss, and the *audiometrist* deals with measures of the loss of hearing, through a variety of audiologic tests. One of the most common tests is the *pure tones audiometry test*, where the listener is stimulated with sinusoidal tones of adjustable frequencies and levels. These tones are presented to the listener through a headphone (*air-conduction test*) or through the bone located behind the ears *(bone-conduction test*). By stimulating the listener with different tones, it is possible to draw an *audiogram*, which gives indications about the degree of deafness. Another type of test is named *speech audiometry* and uses pre-recorded words and phrases, presented to the listener with different levels of intensity. Using these tests it is possible to establish the category of deafness: *conductive* or *sensoneural* deafness. Conductive deafness is caused by some form of impediment in the sound transmission in the inner ear; sensoneural deafness is caused by problems in the inner ear, or in the first part of the acoustic nerve that connects the inner ear with the brain.

Conductive deafness

Conductive deafness may be caused by the following reasons: the outer ear can be inflamed because of an infection (*otitis externa*); the tympanic membrane may be damaged or perforated by some

external body, or by an extremely strong noise, such as an explosion. More often the middle ear is affected by problems. An acute inflammation of the middle ear (*acute otitis media*) caused by an infection in the respiratory tract can become chronic (*chronic otitis media*) increasing the risk of serious damage. An obstruction in the Eustachian tube may lead to an infection which causes the medium ear to be filled with fluid, that becomes more and more viscous with time (*secretory otitis media*). In other cases a nodule formed by spongy bones develops at the basis of the stirrup bone (*otosclerosis*) and reduces the mobility of the stirrup bone, so reducing its capacity to transmit sound to the inner ear. In some case the otosclerosis expands until it reaches the inner ear, causing a gradual degradation of the acoustic nerve.

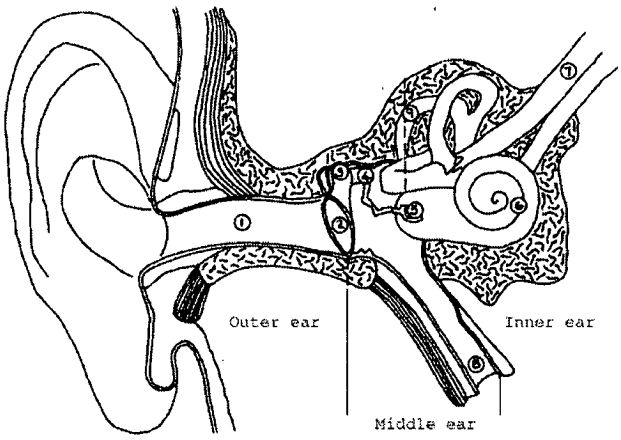


Figure 2.18. Structure of the human ear. 1:Ear Canal (external auditory meatus) 2:Eardrum (tympanic membrane) 3:Hammer Bone (malleus) 4:Anvil Bone (incus) 5:Stirrup (stapes) 6:Cochlea 7:Auditory Nerve 8:Eustachian Tube 9:Semicircular Canals (from McCracken & Sutherland, 1991)

Sensoneural deafness

Sensoneural deafness often involves the cochlea, which is sensitive to damage for many reasons. Illnesses such as rubella, meningitis, or mumps can have direct consequences for the cochlea, and so can the excessive use of certain drugs, such as aspirin. Strong and prolonged noises can reduce the cochlea's efficiency. Other possible causes are congenital malformations in the inner ear. Generally there is little that can be done to solve these problems. In the case of profoundly deaf people, it is possible now to use a recently developed technique, the *cochlear implant*, where thin wires implanted in the cochlea electrically stimulate the cochlea with signals coming from a hearing aid. These stimuli give acoustic sensations and patients have to learn how to decode these new stimuli. In rare circumstances an individual may contract an illness, typically a tumour, in the first part of the acoustic nerve (*acoustic neuroma*). Finally *Meniere's disease* affects the labyrinth in the inner ear, and involves both the vestibular nerve, causing fits of dizziness, buzzing and roar in the ears, and the

auditory nerve, with reductions of hearing capabilities. The cause of the disease is an excessive endolinphatic fluid in the inner ear.

An important issue in the diagnosis of the type of hearing impairments is the ability of the patient to perceive in different ways the sound stimuli in the air-conduction test and the bone-conduction test. If the patient cannot hear sounds through the air, but can perceive differences between these two modes, the deafness is probably neural.

In addition to the causes described above, there may be causes of a mixed nature, where both conductive and neuronal deafness are present at the same time. An example of that is the normal decrease in hearing capacity in older people (*presbycusis*), especially in the high frequency range. Speech is perceived with missing parts (especially the consonants) and the listener needs more time to decode the signal. This is often the cause of the delayed response that some older people show in answering, which is sometimes erroneously confused with a reduction of the intellectual capabilities of the listener.

Central deafness

Conductive and sensoneural deafness are generally grouped under the term *peripheral deafness*, while the term *central deafness* refers to the loss of hearing due to damage of part of the hearing nerve that connects with the brain, or damage in the hearing centres of the brain. In these cases it can happen that the hearing system appears to be normal with pure-tone audiometric tests, but that the patient is not able to interpret the sounds of the speech. The issue in diagnosis is to understand if there are auditory problems, or if there is some sort of linguistic pathology instead, or other deep-rooted problems (e.g. autism). In the phenomenon called *acoustic agnosy* the patient may have difficulties in identifying, as equal, two repetitions of the same word or sound. In *pure word deafness* there is the ability to recognise isolated sounds, but not to integrate these sounds into meaningful units; this syndrome is sometimes named as *auditory verbal agnosis*. In *Wernicke aphasia* there is the ability to recognise and integrate sounds, but the inability to understand them.

Linguistic pathologies resulting from brain damage are mostly caused by cerebro-vascular accidents (CVAs), generally called *strokes*. They are caused by damage to the cerebral cells as a result of a reduced oxygen supply. There are various reasons for this, the most common being arteriosclerosis, where the arteries tend to fill with deposits of fat which not only reduce the useful dimension of the arteries, but also can cause an obstruction by clots of blood. In other cases blood pressure can decrease too much, such as in the case of light heart attacks, and the blood stream can be inadequate to supply the brain cells. Another possibility is an embolism, where a blood clot or a extraneous body enters the blood stream, and blocks the cerebral arteries. An artery can have small leakage in some parts of its walls (*aneurysm)* which can cause an haemorrhage. In such cases generally there is

diffused damage that affects more than one cerebral function. The brain may be also damaged by causes that are different from circulatory malfunctioning. Examples are tumours and many types of infections, in addition to damage from traumas. Nevertheless, the cases in which deafness is caused by brain damage are rare: in most cases it is the result of illness, or damage in the outer, middle or inner ear.

2.4 Speech Production of Hearing-Impaired Talkers

A wide range of studies have been undertaken to characterise the speech of deaf or hearing-impaired people. This section will present the most important results of these studies.

The characteristics of the speech of people born with hearing problems, or people who have developed hearing problems before the age of 1 or 2 years (*pre-lingual deafness*) may be rather different from characteristics of the speech of people that acquired hearing problems in childhood or adulthood, after using spoken language for some time (*post-lingual deafness*).

Characteristics of the speech of pre-lingual deaf speakers have been discussed in many articles and studies over the last fifty years. One of the most important articles is the one from Osberger and McGarr (1982), that deals mostly with the voice of profoundly congenitally deaf children. In contrast, relatively few studies have been carried out on the speech of post-lingually deaf people. From the studies that have been reported however, it appears that post-lingually deaf speakers make the same types of errors as pre-lingually deaf speakers, with the exception of intonation problems. Furthermore, the severity of these errors is generally much less than that for pre-lingually deaf people.

In the following discussion, when appropriate, the distinction between pre- and post-lingually hearing-impaired speech is specified.

2.4.1 Intensity

The control of vocal intensity is a common problem in hearing-impaired speech, especially in the cases of neurosensorial deafness (Penn, 1965). Talkers may have too loud or too soft a voice, and the intensity may vary in an unpredictable way (Martony, 1968). An interesting feature of some hearing-impaired talkers is their attempt to use intensity to communicate intonation, instead of using changes in fundamental frequency as is the case for normal speech (Phillips et al., 1968).

The problems of intensity are similar for both post-lingually and pre-lingually deaf. Leder at al. (1987c) report that hearing-impaired talkers use higher vocal intensity levels overall. Nevertheless

other studies report that post-lingually deaf talkers do maintain the normal use of intensity for signalling the starts and ends of breath groups and of phrases, independent of the age when the deafness was established (Binnie et al., 1982; Plant, 1983; Plant & Hammarberg, 1983).

2.4.2 Fundamental frequency

Fundamental frequency level

To the lay person, the abnormal changes in fundamental frequency exhibited by hearing-impaired talkers are often judged as the primary contributor to the voice quality named "deaf voice". Many studies have reported that adult deaf people usually show a value of fundamental frequency which is measurably higher than normal-hearing adults, for example Angelocci et al. (1964), Boone (1966), Martony (1968), Gilbert & Campbell (1980). Acoustic studies with children's speech found no significant differences in children 6 and 12 years of age (Boone, 1966; Green, 1956; Monsen, 1979), but some differences were found in older children (Boone, 1966). Osberger (1981) whose results for a group of deaf girls, of age between 13 and 15, showed an average fundamental frequency which was some 75 Hz higher than normal. In a study by McGarr & Osberger (1978) perceptive estimates on 50 children of 10 and 11 years showed little perceived difference from the normal in fundamental frequency, but in these studies no attempt was made to assess the levels of variation in the fundamental frequency that can be tolerated before it is considered deviant.

An increased level of fundamental frequency was found in the speech of post-lingually deaf people (Leder et al., 1987a, Lane & Webster, 1991), but this is not a general characteristic. The speakers in a study by Waldstein (1990) did not show consistent fundamental frequency differences between groups of deaf and normal-hearing people. Intrinsic deviations of the fundamental frequency of vowels are also reported by Lane (1988).

Fundamental frequency changes

Another important problem regarding fundamental frequency (and one which makes it difficult to estimate an average level of reliability) is the presence of excessive and inappropriate changes in the fundamental frequency of some speakers (Monsen, 1979; Smith, 1975b; Stevens et al., 1978). These changes are sometimes accompanied by a complete loss of phonatory control, and by the absence of phonation. Speakers can sometimes change to *falsetto* (a different mode of vibration of the vocal folds). In contrast, some speakers can show a reduced range and variation of fundamental frequency, resulting in a monotonous voice (Angelocci, 1962; Monsen, 1979). Bush (1981) suggests that some of the problems of fundamental frequency in deaf people, particularly the occurrence raised fundamental frequency on stressed syllables, may be caused by an excessive tension of the larynx. Lane (1988)

also found evidence of an increased variability of fundamental frequency in the speech of the post-lingually deaf. This variability of fundamental frequency seems to be particularly high in stressed vowels (Lane & Webster, 1991).

The biggest difference between post-lingually and pre-lingually deaf speakers is in the use of fundamental frequency for intonational purposes. In Waldstein's (1990) study, for example, many speakers maintained a distinction between the pitch declination in sentences, and the pitch slopes in polar questions; speakers that did not have this distinction were the ones who had lost their hearing early in their lives.

2.4.3 Vowels and diphthongs

Vowels

Errors in the production of vowels are reported in many studies on pre-lingually deaf (for example Hudgins & Numbers, 1942; Geffner, 1980; Markides, 1970; Smith, 1975b). According to Osberger and McGarr, there are generally fewer errors perceived for vowels than for consonants. The reason for this can be attributed to the fact that people evaluating the speech of deaf people tend to tolerate a greater distortion for vowels than for consonants, before reporting an error. The reported errors are variously described as substitution, neutralisation or centralisation errors. Deaf speakers tend to have a reduced contrast in the vowels they produce, with articulations always more central than they should be, that is they have a reduced vowel frequency space. These vowels sound more neutral and indistinct, and they can greatly vary in quality between phrases.

It is reported that back vowels are more often correctly pronounced than high or middle vowels (Geffner, 1980; Nober, 1967; Smith, 1975b). Other errors involve the confusion in tense-lax vowel pairs, such as /i/ (as in "bead") and /I/ (as in "bid") (Smith, 1975b) and the diphthongisation of pure vowels (Boone, 1966; Markides, 1970; Smith, 1975b). The neutralisation of vocal contrast has been observed acoustically in studies by Angelocci et al., (1964b), Monsen (1976a, 1978), Osberger, Levitt & Slosberg (1979), and, for French speakers, by Perrin et al. (1994).

Many studies have reported hearing-impaired speakers having limitations in the movements of their tongue. According to Boone (1966) for example, deaf speakers tend to maintain the tongue in a back, low position. This seems to match the results of acoustic studies like those by Monsen (1976a), who observed the relative immobility of the second formant F_2 of the vowels at around 1800 Hz, which is above the residual hearing capability of most deaf people. These studies suggest that the immobility problems of the speakers refer particularly to the front-back dimension of the tongue movements, an effect that is more difficult to observe (and probably to learn) than the more visible high-low dimension.

As regards post-lingually deaf speech, errors in the production of vowels are reported by Cowie & Douglas-Cowie (1983), particularly the centralisation of vowels "along the acute axis", so that the distinctions between /i/ and /I/ (as in "bead" and "bid") , /e/ and /a/ (as in "bed" and "bad"), and /a/ e /ɑ/ (as in "bad" and "bard") are blurred. A limited distribution of vocal segments is also reported by Waldstein (1990).

Diphthongs

Errors were also found in production of diphthongs, although these are not so extensively covered in the literature as errors in the production of vowels. Typical errors include the simplification of diphthongs, with the omission of one or the other vocalic element (Hudgins & Numbers, 1942). Very few acoustic studies have been carried out on this phenomenon, apart from a detailed study of the diphthong /aI/ (as in "buy") by Monsen (1976d). This study found a variety of patterns in the speech of deaf people, one of which exhibits a large variation in the frequency of the first formant F_1, while the second formant F_2 remained almost unchanged. Monsen explained this phenomenon suggesting that the speaker was moving the jaws correctly, but without the corresponding correct movement of the tongue. This reinforces the theory of a limited mobility of the tongue in studies previously quoted.

2.4.4 Consonants

Voicing errors

One of the most common errors in the speech of pre-lingually deaf people is the confusion between voiced and un-voiced consonants. Voiced stops are replaced by the unvoiced correspondent, and vice-versa (Hudgin & Numbers, 1942), however there are contrasting opinions about the direction. In studies by Smith (1975) and Heider et al. (1941), speakers substituted voiced sounds with non-voiced sounds, while Markides (1970) found that most errors involved the substitution of non-voiced sounds with voiced sounds.

The difficulty in characterising this type of error lies partially in the fact that many studies used judgements based on author's impressions. Physiological studies (for example McGarr & Løfqvist, 1982; Whitehead & Barefoot, 1980) suggest that vocalisation errors are not the simple substitution of a class of sounds with another one, but they are the result of the inability to co-ordinate the timings of breathing, phonation and articulation. This may reduce the difference between the voiced and non-voiced sounds, and make it very difficult for the listeners to characterise the perceived sounds. This is confirmed by acoustic studies on voiced/non-voiced errors, which show a reduction in the difference between the acoustic parameters of the voice onset time measured in voiced and non-voiced stops (for example Monsen, 1976b; Mahshie, 1980; McGarr & Løfqvist, 1982). In any mode, other acoustic

parameters can also be cause of this confusion. One factor is segmental duration (discussed later in section 2.4.6): deaf speakers tend not to differentiate the duration of the stop closure between voiced and non-voiced stops (Calvert, 1962), unlike normal-hearing speakers, and this produces a longer closure duration for non-voiced stops. Furthermore, many studies have shown that differences in the duration of the preceding vowels, which help in identifying vocalisation in stop and fricatives, may be absent in the voice of deaf people (Calvert, 1961; Monsen, 1974).

Voicing errors made by post-lingually deaf speakers appear to be of the same type as those made by speakers who were born deaf. According to Cowie & Douglas-Cowie (1983), many errors involve the substitution of vocalised segments with non-vocalised plosives and labio-dental fricatives. In a study on seven post-lingually deaf speakers in order to understand the effect of the loss of auditory feedback on some speech related parameters, Waldstein (1990) found that the voice onset time of non-voiced stops were shorter for deaf speakers than for a control group of normal talkers (with matched age and sex), even if differences in voiced stops were not acoustically identifiable.

Place of articulation errors

Place of articulation errors are common in the speech of pre-lingually deaf people. With these errors, a consonant is replaced by another consonant produced with a different place of articulation. Consonants produced towards the back of the oral cavity, for example palato-alveloar, palatal and velar sounds, are particularly susceptible to being replaced (Geffner & Freeman, 1980; Levitt et al., 1976; Geffner, 1980), by either a frontal consonant or a glottal stop (Smith, 1975b; Levitt et al., 1976). An explanation often given for this fact is that the sounds produced toward the front of the oral cavity, employing the tip of the tongue or the lips, are more visible, and then more easily acquired by the deaf talker in the absence of acoustic feedback. Osberger & McGarr (1982) suggest that the physical limitations of some articulators could have an effect, since articulators such as the lips have a narrower range of movement, and this gives less possibility for positioning errors.

Acoustic studies on the production of consonants in deaf speech suggest that the transitions of the formants between consonants and vowels, which give the most information about the place of articulation in normal speech, are partially or completely missing in the speech of deaf people (Martony, 1965; Monsen, 1976c; Rothman, 1976). Furthermore, these formant transitions change little with the change of phonetic context, and also many of the clues used by the listeners to identify the place of articulation are missing.

Errors in place of articulation are also typical in the post-lingually deaf, apparently caused by problems in fine control of the position of articulators. The articulators for fricatives seem particularly liable to this type of error. Lane (1988), for example found that the acoustic distinction between /s/ (as in "sea") and /ʃ/ (as in "ship") was reduced in post-lingually deaf speakers, even if still present. A

study from Lane & Webster (1991), examining aspects of the voice of three post-lingually deaf speakers, found that the spectral centres of /s/ and /ʃ/ were in fact very close together. Other studies (Tartter et al., 1989; Cowie & Douglas-Cowie, 1983) suggest that the cause is the more frontal articulation of the /ʃ/ in these speakers. The study by Lane & Webster (1991) also suggested that the spectral information which marks the place of articulation in stop bursts is not so well differentiated in the speech of these deaf people.

Manner of articulation errors

Errors of manner of articulation imply the substitution of a segment type with another, produced in the same place of articulation, but using a different mode. Errors of this type commonly seen in the speech of deaf people include the use of stops instead of fricatives, and the use of stops or fricatives instead of affricates. Labial stops, glides like /w/ and /r/ (as in "way" and "ray"), and the labia-dental fricatives /f/ and /v/ (as in "fan" and "van") are often correctly produced (Smith, 1975b). Osberger & McGarr (1982) suggest that these errors result when speakers correctly position their articulators in terms of place of articulation, but fail in the co-ordination of the movements required for the correct degree of *stricture* (distance between articulators).

Omission errors

Errors involving the omission of consonants appear to be the most common among all reported errors (Hudgins & Numbers, 1942; Markides, 1970; Smith, 1975b). Typical errors include the omission of a consonant at the beginning (especially /h, l, r, j, θ, s/, as in "hat", "lay", "ray", "yes", "thin", "sea") and at the end (especially /l, t, s, z, d, g, k/, as in "ball", "hit", "price", "prize", "fade", "dog", "sock"). Hudgins & Numbers found that the production of final consonants can have many forms: consonants were completely omitted, released on the following syllable, or incompletely produced.

Consonant cluster

Hudgin & Numbers (1942) also found a large number of errors in the production of consonant clusters. These were often simplified, with the omission of one or more components (for example, /st/ produced as /t/ as in "stall") or with the insertion of an intrusive vowel (generally schwa) between the components of a cluster (Smith, 1975b). This kind of error, as Osberger & McGarr point out, can be particularly detrimental to perceived timing and rhythm, and can cause considerable difficulty for a listener in understanding the words.

2.4.5 Voice Quality

Nasality

The reduced control of nasality is another area where deaf people have significant problems. There are many descriptions of the speech of deaf speakers as "nasal", especially in older studies (for example Hudgins, 1934), but more recent studies on speech production (acoustic and physiological) have not always given evidence that the reason for nasal voice is a lack of control in the velopharingeal opening, which is the normal articulator of nasality. There is no clear correlation between perceived nasality in the voice of deaf speakers and the velopharyngeal opening or the air stream (Seaver et al., 1980; Stevens et al., 1983), and it seems that other causes, such as the limited control of articulation, of timing and of fundamental frequency, may partially influence the listener's judgement (Colton & Cooker, 1968).

Some problems in the coordination of the velopharingeal opening were found in a study by Stevens et al. (1976). They measured nasal resonance using an accelerometer, to overtake the problem of perceptive judgement of nasality, and they found that in fact speakers had problems in co-ordinating velopharyngeal movements, particularly in nasal stop clusters. Stevens at al. (1983) suggest that inappropriate timing in the opening and closing movements of the velum can lead to the increased perception of nasal resonance.

Phonation

Pre-lingually deaf speakers often have great difficulties in controlling the mode of vibration of the vocal folds to produce efficient phonation. The resultant speech is then characterised by increased harshness, often caused by excessive laryngeal tension (Whitehead & Emanuel, 1974; Wirz et al., 1979), or breathy phonation, as a result of an inefficient vibration of the vocal folds, and with insufficient closure. Breathy phonation causes an excessive use of air, and this may be a reason why deaf speakers often are short of breath before they manage to end a sentence. Other problems reported in this area include the use of diplophonia, or creaky voice (Monsen, Engebretson & Vemula, 1979), and problems in co-ordinating the necessary laryngeal movements for voiced and unvoiced segments (Metz, Whitehead & Mahshie, 1982).

Post-lingually deaf speakers appear to have fewer problems in controlling their phonation, although some studies report an increased level of breathiness in vowels, (for example Lane, 1988). Interestingly Waldstein (1990) reports reduced levels of pitch jitter (the standard deviation of the duration between cycles, normalised against the fundamental frequency) in deaf speakers. According

to Waldstein, reduced jitter levels appear to be common in post-lingual deafness, even though the reason is not clear.

2.4.6 Timing

Speech rate

It is commonly reported that deaf speakers speak much more slowly than normal-hearing speakers. Boone (1966) for example, found that pre-lingually deaf speakers take 1.5 to 2 times longer than hearing people to say the same sentence. There appears to be two reasons for this reduced speed. Firstly there is a tendency to make segments longer, especially vowels (Osberger & Levitt, 1979). Secondly there appears to be an inappropriate use of pausing (Boothroyd et al., 1974; Boone, 1966). A reduction in the speech rate in post-lingually deaf, with a consequent increase of the time necessary to speak, is reported by Lane & Webster (1991).

Speech segment duration

Studies on the timing of the speech of deaf people have shown that excessive prolonging of segments is one of the most important problems. Vowels, fricatives, and the closure part of stops can have a duration which is up to five times longer than normal (Calvert, 1961). According to Osberger & Levitt (1979), the biggest difference between normal speech and the speech of deaf people can be attributed to lengthening of vowels. However, Osberger & McGarr (1982) underline that comparisons of segment duration are complicated by the fact that deaf speakers can shorten some syllables by omitting some segments.

It was also observed that deaf speakers show a greater variability in the control of the duration of segments in comparison with normal-hearing speakers (Osberger, 1978). They also are unable to produce many of the temporal patterns used in normal speech, such as the use of varying the length of vowels for signalling contrastive stress (McGarr & Harris, 1980); the use of varying the length of vowels to emphasise the voiced - unvoiced distinction in following obstruents (Calvert, 1961; Monsen, 1974); and the distinction in the length of articulatory closures that are normally observed in voiced and non-voiced stops (Calvert, 1962). Studies on the use of duration to signal the structure of pauses in speech have shown that some deaf speakers retain this control (Reilly, 1979) while others do not (Stevens at al., 1978).

Syllables, phrases and paragraphs which are longer than normal have been reported in the speech of post-lingually deaf (Leder et al., 1987b; Lane, 1988). Part of this increase seems to be caused by the increased movement of the articulators (Zimmerman & Rettaliata, 1981). Waldstein (1990) also found an increased variability in the duration of segments in deaf people. However, in this study the

speakers maintained the length contrast of vowels which cues the voiced status of the subsequent stops (as compared to pre-lingually deaf speakers).

2.4.7 Pausing, Breath control, Rhythm

Pausing

Deaf speakers have noticeable problems in controlling pauses. They insert a large number of pauses in inappropriate positions (Boothroyd, Nickerson & Stevens, 1974; Boone, 1966), and often, as a result produce a large number of short and irregular breath groups (Hudgins & Numbers, 1942). According to Stevens at al. (1978), the length of pauses constitutes the biggest difference in the characteristic of timing between normal speech and deaf speech.

Breath control

It is often reported that deaf speakers cannot co-ordinate the breathing, phonatory and articulatory processes used in voice production. An aspect of that is the co-ordination between breath control for speech production and the needs of respiration: speakers often start speaking with insufficient air volume in their lungs, or they waste air while they speak because of an inappropriate opening of the larynx, or because of inefficient phonation (McGarr & Løfqvist, 1982). This results in the air supply being exhausted before the end of the utterance (Forner & Hixon, 1977; Whitehead, 1983), so reducing the length of the utterance itself.

Rhythm

The marking of stress is one of the rhythmic characteristics that can be affected in the speech of deaf people. Deaf speakers often distort stress patterns in utterances by stressing the wrong syllable, or wrongly controlling parameters such as fundamental frequency, intensity or vowel quality, so that any syllable may result as stressed. Boothroyd, Nickerson & Stevens (1974) for example, found that deaf speakers lengthen non-stressed syllables, so reducing the contrast between stressed and non-stressed syllables. Furthermore, the frequent use of pauses in the wrong positions, the insertion of vowels in consonant clusters, and the complete omission of some segments and syllables, make it extremely difficult for a listener to reconstruct the rhythmic pattern of an utterance from such a deaf talker.

2.4.8 Perceptual analysis of hearing-impaired speech

A different approach in the characterisation of the speech of deaf people was attempted by Wirz (1987), using a perceptive labelling technique called Vocal Profile Analysis (Laver at al., 1981). This technique tries to characterise the speech analysing a wide range of speech settings, instead of analysing the production of single sounds. These settings include the normal use of the speaker's phonation, and any long term variation of tongue, lips and jaw position.

In Wirz's study, vocal examples from 50 moderately or profoundly deaf people were evaluated by four experts using the Vocal Profile Analysis method, and characteristic speech settings were derived. The derived characteristics were:

- reduced extension of the movements of tongue, lips and jaw;
- increased tension in the laryngeal, pharyngeal and supra-laryngeal area;
- increased harshness and increased creakiness in the voice;
- raised position of the larynx, often accompanied by pharyngeal constrictions and by retraction of the tongue position (backed tongue body);
- reduced pitch range, with a reduced variability;
- abnormally perceived pitch, too high or too low;
- reduced extension of vocal intensity, generally too low.

According to Wirz, the main setting which characterises deaf speech is a general increase of the vocal tract tension. Wirz asserts that this setting may be adopted by deaf speakers in the attempt to increase the available kinestatic feedback to improve self monitoring. The adoption of this setting has ramifications in all aspects of voice production, including phonation (since an increased tension causes harshness in the voice and a lack of control of fundamental frequency), articulation (because of a restriction in the movements of the tongue) and timing (since vowels may be lengthened, but with rapid transitions from vowel to consonantal articulations). Wirz suggests that increased laryngeal tension may also be the cause of the perception of an increased pitch which is observed in deaf speakers, although this is not evidenced from measurements of the fundamental frequency: one of the spectral characteristics of tense voice is an increase of the energy in the middle frequencies (Laver, 1980), which may give the perception of an increased pitch. According to Wirz, speech therapy for deaf people should concentrate on teaching speakers how to reduce the overall tension; Wirz also suggests that a way to achieve this result is to eliminate the dependency on kinestatic feedback through visual feedback.

2.5 Conclusion

This chapter presented the process of normal speech production, and showed how this process can be affected by deafness. The lack of auditory feedback, which breaks the closed loop system known as the speech chain, can partly be substituted with alternative forms of feedback, such as tactile or visual feedback, the latter being the topic of this thesis. Chapter 3 presents a review of visual feedback techniques used in the rehabilitation of hearing-impaired speech. However these techniques are not always adequate, and they present a range of problems. An analysis of these problems, and a novel approach for reducing them will be presented in the following chapters.

CHAPTER 3

A Review of Visual Feedback Techniques in the Rehabilitation of Hearing-Impaired Speech

3.1 Introduction

The previous Chapter showed how deaf people have difficulties in developing normal language skills because of their lack of vocal feedback. This feedback can however be substituted in more than one way, for example with vibro-tactile devices, or with visual feedback. It is not within the scope of this thesis to address these two methods. Nevertheless it is important to note that whilst vibrotactile devices may provide perceptions of some aspects of voice, for instance the voice level and, to some extent the fundamental frequency, visual feedback can only create perceptual metaphors that the user has to interpret. The skin is able to perceive vibrations which provide information about the intensity and frequency of a sound. Better results are however obtained by frequency scaling, and optionally by redundant coding of F_0 as both frequency and spatial position on the skin (Bernstein, 1995). In contrast the eye can perceive repetition rates to around 10 Hz and, as discussed in Chapter 4, studies on flash rates report that it is quite difficult for the eye to differentiate more than five different flash rates (Grether & Baker, 1972). The goal is therefore to find alternatives, metaphors in fact, for providing a visual substitute of the voice. This Chapter reviews studies in the field, and gives examples of visual feedback used in published systems.

3.2 Previous work in the field

This section reviews the main studies on research of visual interfaces as an aid for deaf voice rehabilitation published during the last 20-30 years. Research in this field forms part of the more general developments in speech training systems for a variety of speech disorders. These include articulatory disorders such as with cerebral palsy patients, and patients suffering of dyspraxia (stroke victims, for example), and speakers with phonological rather than physiological problems. These patients have some speech problems in common with deaf speakers, such as lack of control of pitch, loudness and articulation, but many have the advantage of retaining auditory feedback, therefore they are not completely dependent on visual feedback. This review deals only with systems specifically designed for the deaf, and takes these limitations of auditory feedback into account.

3.2.1 User groups

As shown in the previous Chapter, the effect of deafness on speech varies enormously, and the needs of the different users have to be carefully considered in the design of any speech training aid. These needs impact most aspects of system design, from the nature of the user interface to the speech material used in the courseware. The age of the user has to be taken into account, since this greatly influences the choice of appropriate display formats for speech feedback.

For example, many speech training systems are especially designed for children, some as young as three or four years old. Many studies (for example see Brooks et al., 1981; Arends et al., 1991) reported that children benefit from computer-based speech training systems more than other user groups, and it seems that the earlier the training starts, the better are the benefits obtained. This probably depends on the fact that young children are still in a very active stage of language development, and they are able to process the visual information from the rehabilitation system in a way that resembles the way that hearing children use auditory feedback. The development of a computer system that can be used effectively by children is often challenging, since children are subject to a number of limitations which adult users generally lack. For example, children have less well developed language capabilities, with a much more restricted vocabulary, limiting the kind of language which can be used in both instruction and practice materials. Furthermore, children have limited ability to concentrate, and become bored and lose interest quickly. In the case where the computer has to be used directly by the child, the limited manual dexterity typical of young children must be taken into account (Javkin et al., 1993a). It is clear that a graphical display capable of attracting a child's attention, and motivating them to interact with the system will result in more time spent in front of the system using their voice, and is a good pre-requisite for a successful therapy. To

motivate the youngest users, it has been suggested that instructional programmes (not only in the speech rehabilitation field) can be made more interesting, and more enjoyable, using the same characteristics that contribute to the "compellingness" of video arcade games (Malone & Lepper, 1983). These characteristics include explicit goals; sound effects; elements of causality; use of graphics instead of words to give instructions and feedback; scoring (Malone, 1981). Displays for deaf speech rehabilitation can be designed to include many of these characteristics, except obviously, sound effects.

In contrast, adults may prefer alternatives to videogame-type displays, and will be more interested in more qualitatively and quantitatively explicit aspects, such as pitch frequency and loudness range, and spectrographic displays. As discussed in Chapter 4, particular care has to be taken in order to avoid de-motivating hearing-impaired adults by showing them the results of their speech capabilities in comparison with normal speaker's capabilities as a performance target..

3.2.2 Physical basis of feedback

Studies of speech production (Flege, Fletcher & Homiedan, 1988; Gay, Lindblom & Lubker, 1981) have shown that the same perceptually corrected sound can be produced with a considerable range of different articulatory gestures, or movements. This suggests that feedback should be given on the basis of the waveform, rather than on the mechanical action of articulators or electrophysiological responses associated with them. In contrast, other research results have shown that stable control of the position of the articulators which are normally not externally visible can be obtained through visual feedback of articulator positions (Fletcher & Hasegawa, 1983; Fletcher, McCutcheon, Martin & Smith, 1988; Mahshie, Alquist-Vari, Waddy-Smith & Bernstein, 1988; Yamada, Murata & Oka, 1988). Research with physiologically based feedback showed that a level better than a coarse motor co-ordination of speech events cannot be achieved. The tendency is to use the speech waveform as the basis for feedback, for the relative simplicity of extraction, and for the proven validity of the speech waveform as the final judgement criteria on the quality of a production. In some studies, however (for example Rossiter & Howard, 1994), the waveform is used to show real-time animation of the approximation of the area of the vocal tract.

3.2.3 Amount and type of information displayed as feedback

Watson & Kewley (1989a) suggest that the display designed for teacher-guided training sessions needs to be simple and also requires an interpretation by the therapist to give feedback to the user. Displays designed for autonomous use, without continuous supervision by a therapist, instead require more advanced programming techniques to give the user a clear meaning of the feedback.

According to Povel & Maassen (1987), the fundamental questions to be answered in order to develop an effective visual feedback are *what* to show and *how* to show it. Attempts to answer these questions have led to fundamental investigations of perception and speech production, as well as basic questions about the essential differences between hearing related processes, and vision related processes. In deciding to ignore incomplete understanding of these complex processes, and to follow a more practical approach, Povel & Maassen drew a set of basic assumptions:

- There must be a unique and fixed relation between acoustic parameters and visual correlates. This implies that the same graphic cannot be used to show different acoustic parameters. Furthermore, if an acoustic characteristic is associated with some visual attribute, this should not be changed later.
- Visual information should be as complete as possible, independent of the fact that the user may or may not use all of the information at the same time. For example, working with a display representing pitch, the voice level should be always visible, to avoid the situation where the target pitch is produced at excessive and harmful voice levels, or with bad voice quality.
- Since the display shows a great deal of information simultaneusly, these should be shown in an integral way, instead of in a parallel way (for example in different windows in the screen). This may be realised using different independent visual dimensions, such as shape, colour, texture and dimension.

In another study, Watson & Kewley (1989) created a series of guidelines, taking into account hints from the fields of human engineering, learning and human cognition and, in the case of children, developmental psychology:

- Salience. The critical information necessary to improve erroneous speech production should be presented as the salient characteristic of the display, and not hidden in a quantity of details of less importance.
- Dynamic tracking. At the beginning of the rehabilitation of articulator movements, feedback based on an accurate production of the steady properties may be sufficient. In the rehabilitation of syllables, words and phrases, however, feedback may have to involve the dynamic movements of the articulators.
- Delayed feedback. When basic levels of articulation control is achieved, the rehabilitation shifts to the word level, and the feedback should be presented *after* the production is completed, but with a delay not exceeding 500 ms. Longer delays reduce the efficiency of the system, and reduce the willingness of children to attend sessions.
- Mono-dimensional feedback. The amount of information shown depends on what is being taught. At the beginning of rehabilitation - or even later when say, a particular characteristic of production

seems to be lost - a single-dimensional display may be useful. Such feedback may be shown in real-time, or at the end of voice production. A problem with single-dimensional feedback is that the user is not aware of possible worsening in the production of other dimensions.

- Multi-dimensional feedback. In comparison with mono-dimensional feedback, the benefit is that the user can be alerted to possible deterioration in other dimensions. It is clear, however, that complex multi-dimensional displays such as spectrograms are inappropriate for rapid interpretation (Cole & Zue, 1979), especially for use with children. In general, the choice of a particular combination of feedback dimensions is a compromise between the maximum quantity of data that the user can process, and the minimum necessary quantity of information for effective training.
- Evaluative feedback. When the user has learned how to produce a sound of acceptable quality, prolonged drill is generally necessary before the intelligibility level can be achieved *automatically*[1]. An evaluative feedback able to determine the quality of a production is particularly useful in prolonged drill to reach automaticity. It should be considered, however, that the validity of an evaluation algorithm needs to be established on the basis of a comparison with a human judgement of the same aspect of speech quality as the algorithm is intended to evaluate.

In more recent studies, Arends (1993) synthesised the studies from Powel & Wansink (1986) and considered a range of published systems, such as BBN (Nickerson, Kalilow & Stevens, 1976), Visipitch (De Bot, 1983), STS (Ferguson, Bernstein & Goldstein, 1988), VideoVoice (Stoker, Fitgerald & Gruenwald, 1987), SpeechViewer (Ryalls, 1989; Wempe & Lunen, 1991) and systems to teach voice intonation to normal-hearing people (Spaai, 1993). Arends defines guidelines for the display of various aspects of the voice, as follows:

- Mapping of a basic speech aspect onto an available visual dimension should satisfy the following criteria: its appearance must be salient, unique, natural, simple, and easily interpretable.
- Suitable visuo-acoustic relations are: colour (voicing), size and brightness (loudness), vertical location (pitch), horizontal location and movement (time), plane position (vowels/consonants), intrinsic object information (nasality detection), indicator arrows (voice quality).
- The display types to be used should be based on either single or multi-dimensional representations, with a preference for integral displays.

[1] *Automaticity* in cognitive research and theory (Shiffrin & Schneider, 1977) is conceived as a perception mode where a large quantity of information can be processed without degrading the performance of other cognitive operations taking place concurrently. It seems probable that fluent speech requires an acquired neuromotor automaticity, in which patterned motor responses are executed without the speaker having to take care of details in production.

The recommendations made by Arends are based on the sometimes contrasting guidelines shown previously, and add a definition of metaphors suitable for rehabilitating various aspects of speech. This definition is based on the principle that an object is characterised perceptually by properties such as shape, colour, brightness, position, angle of inclination and texture. Arends proposes that since these properties are perceived and coded independently in early vision, each one of these dimensions can be used, with due care in considering the natural similarity between the auditory and visual dimensions. Furthermore, *time* may be introduced as a dimension orthogonal to all other speech dimensions, choosing a movement from left to right as the most obvious. (In section 4.2 some critique of these consideration will be presented). Other studies (for example Ruoss et al., 1988) have examined optimal methods to show information visually to normal-hearing subjects, and a new study that is taking place in Gallaudet University for the Deaf (Washington, DC) is investigating the mechanism of visual speech perception both in deaf and normal-hearing people[1].

[1] Spoken Language Processing without Audition (NIH Funded). The objective of this project is to provide an integrated account of the perceptual, linguistic, and cognitive processes underlying visual speech perception and characterise the similarities and differences in those processes among individuals with different types of auditory and linguistic experience. The application cover aid in the speech and language training of hearing-impaired children, speechreading training for adults, and use and development of sensory aids such as digital hearing aids, cochlear implants, and vibrotactile aids.

3.3 Published Systems review

In this section published systems for deaf voice rehabilitation are reviewed, focusing on the visual displays which are used in them.

3.3.1 Type of feedback

Computer based training systems vary in the type of feedback given to the user. At the lowest level a system may give the user some sort of feedback such as pitch contour, or information on vowel locations, without other information about the quality of their production. Systems like this represent a valid support for the speech therapist, since they allow them to focus on a particular aspect, or illustrate a particular point, but they have limited application. At the next level, a system may allow the therapist to compare the speaker's production with some reference or model. Many of the tools that use split-screen displays (for example VisiPitch, De Bot, 1983) are in this category. Information regarding the quality or acceptability of speech production must then be deduced by the therapist or by the client from this comparison. At an even higher level there are *evaluative* systems, which give the user information about the quality or goodness of their attempts through an automatic execution of such a comparison. Evaluative systems represent the majority of systems now under development. The basis of the comparison may be a model stored in the system itself, a model produced by the therapist during a training session or, as in the case of the ISTRA (Watson, C.S., Reed, et al., 1989) and VideoVoice (Stoker, Fitgerald & Gruenwald, 1987) systems, the best attempts produced by the client and recorded with the help of the therapist during a training session. The highest level of sophistication is represented by diagnostic systems, capable of giving feedback on the nature and cause of errors, and advice on how to correct them. No system capable of giving this level of feedback has yet been developed, even if noticeable progress has been achieved in the development of the system of the Boys' Town Institute (Lipmann and Watson, 1979; Osberger et al., 1981; Watson et al., 1989a).

3.3.2 Nature of feedback

Visual feedback can vary from simple parametric displays, showing the behaviour of a chosen parameter such as pitch over time, to more abstract representations of phonetic or acoustic categories, to sophisticated computer-generated video games controlled by the value of the analysed parameter. Many systems allow a choice of displays for different purposes, including displays for calibration and monitoring by the therapist. For use with children in particular, development is moving away from the

more scientifically-oriented screen formats that have characterised previous systems, towards more motivating video-game type formats.

3.3.3 Coverage of speech features

Almost all aspects of speech are covered by one system or another, including pitch control (particularly pitch range and variability, and more rarely intonation), rhythm, vocalisation control, vowel production, nasalization, consonant articulation, laryngeal quality, amplitude and duration of speech. The more common characteristics are those which are more easily measurable from acoustic waveforms, that means that nasality, voice quality and detailed analysis of consonant articulations - all aspects that require specific physiological sensors to obtain the best measurement - are less well represented. The trend in the development of speech rehabilitation systems is moving away from covering one or two parameters as the maximum, such as in the case of the Kay VisiPitch (De Bot, 1983), towards complex systems covering many parameters, like the CISTA (Yamada & Murata, 1991), or the Johns Hopkins Speech Training aid (Ferguson, J.B., Bernstein, L.E., & Goldstein, M.H., 1988), able to give feedback on many parameters at the same time.

3.3.4 Courseware

Systems vary considerably in the way they present the courseware that accompanies the analysis and feedback software. Many systems only give a series of modules or independent games, often with little information about the way to use them. Watson et al. (1989b) underline the importance of correctly prepared and integrated courseware for the success of computer based training, and their ISTRA system is fully integrated in a standard teaching method. In this method, the speaker is first evaluated by the therapist, and a training programme is established taking into account the capabilities of the training system. This is an essential first stage to assure that the use of the aid system is appropriate for the particular speaker. The speaker then follows a progression through a series of training drills, which become more complex as the speaker's capabilities increases. Two systems, the John Hopkins aid and the Visual Speech Apparatus, follow a therapeutic approach underlined by Ling (1976), in which users undertake a series of steps in their development of appropriate patterns in speech production. For example, users of the Visual Speech Apparatus first acquire an awareness of the existence of some dimensions of speech production, then move on to acquire a basic control of these dimensions in their vocal productions, finally proceeding to acquire a more complex and categorical control in a more extended range of linguistic materials (Povel & Arends, 1991; Arends, 1993).

3.3.5 Examples

This section shows examples of visual feedback from several published systems: C-Speech, Dr. Speech Science, IBM SpeechViewer 1 and 2, Kay Elemetrics, Kaway, Panasonic, STS, Visual Speech Apparatus.[1] The examples are grouped in categories.

Vocalisation

In the early stages of speech rehabilitation it is sometimes necessary to make users aware if they are producing any sound at all. After this basic control of the voice is achieved, it is possible to practice more selective characteristics of speech, such as loudness or pitch. In the simplest cases, feedback is designed to simply tell the users when they are phonating and when they are not. They may show an object, for example a car, moving toward the right side of the screen when there is any sound above a certain threshold, and stopping in case of silence. A variation of this idea is the display in Figure 3.1, designed to practice short and repeated vocalisations (for example: "pa pa pa"). Here a new footprint appears every time a sound is produced, getting closer to the target in the right side of the screen.

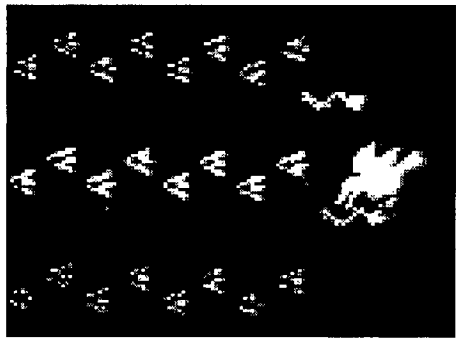


Figure 3.1. "Short and repeated" vocalisation (STS System)

In other displays an object moves constantly from the left to the right of the screen, changing colour depending on the type of phonation. Such displays are often used to practise a sustained vocalisation, in order to improve breath control, or to show the differences between silence, voiced and unvoiced sounds.

[1] C-Speech™, MillgrantWells LTD, P.O. Box 3, Rugby CV21 3UF, England; Dr. Speech Science™, Singular Publishing Group, Inc. 4284 41st Street, San Diego, CA 92105-1197; IBM SpeechViewer™, IBM, 1133 Westchester Avenue, White Plains, NY 10604, United States; Kay Elemetrics Corp., 2 Bridgewater Lane, Lincoln Park, NJ 07035-1488 USA; Kaway, 430, Tokio, Japan; Panasonic, 226, Tokio, Japan; STS (Ferguson, Bernstein & Goldstein, 1988); Visual Speech Apparatus, Instituut voor Doven, Dept Research & Development, Theerestraat 42, 5271 GD Sint-Michielsgestel, The Netherlands.

For example, a rectangle may fill with a colour which varies according to the type of phonation (in the example in Figure 3.2: black = silence, chequered = unvoiced, white = voiced).

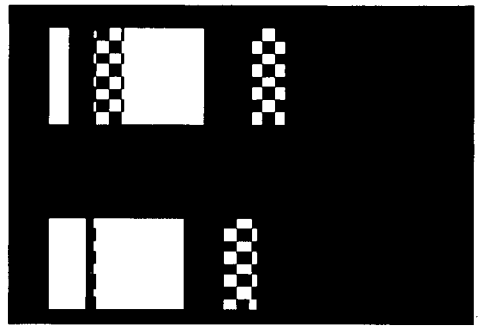


Figure 3.2. "Sustained" vocalisation (C-Speech system)

Intensity

Intensity is represented by an object generally with a round shape (for example a circle or a balloon) which varies in size with the voice level, or a variation of this idea. In the example in Figure 3.3, the height and volume of the water stream from the head of the whale is proportional to the voice intensity.

Figure 3.3. Voice intensity as size of an object (IBM SpeechViewer 2)

Sometimes the brightness or colour of an object can be used, as in the example in Figure 3.4, where the balloon changes size and colour in proportion to the voice intensity.

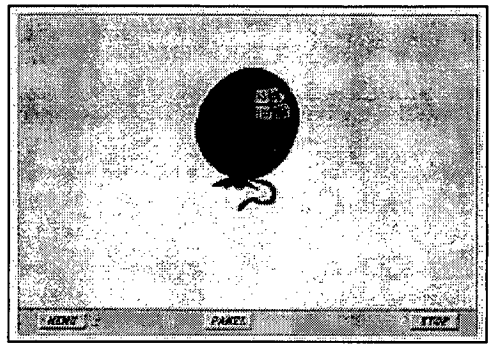


Figure 3.4. Voice intensity as size and colour of an object (Visual Speech Apparatus)

In other systems the voice intensity controls the speed of an object, such as a car, moving from the left to the right side of the screen (such as in the VideoVoice system); in others voice intensity is represented by an object (for example a balloon again) moving up or down with the voice intensity (such as in the STS System).

Fundamental frequency

Fundamental frequency (F_0) is often represented by the vertical position of an object. In the example in Figure 3.5, the neck becomes longer or shorter with the user's F_0.

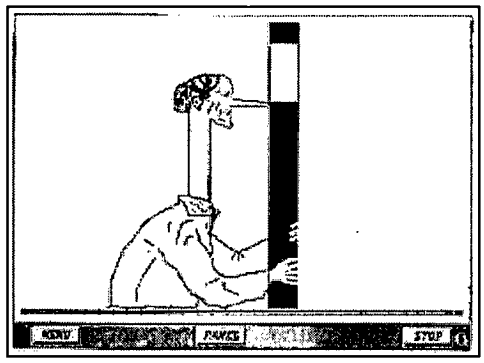


Figure 3.5. Fundamental frequency as vertical position of an object (Visual Speech Apparatus)

Optionally the object moves to the right with time on a frequency/time plane, as in the two examples in Figure 3.6.

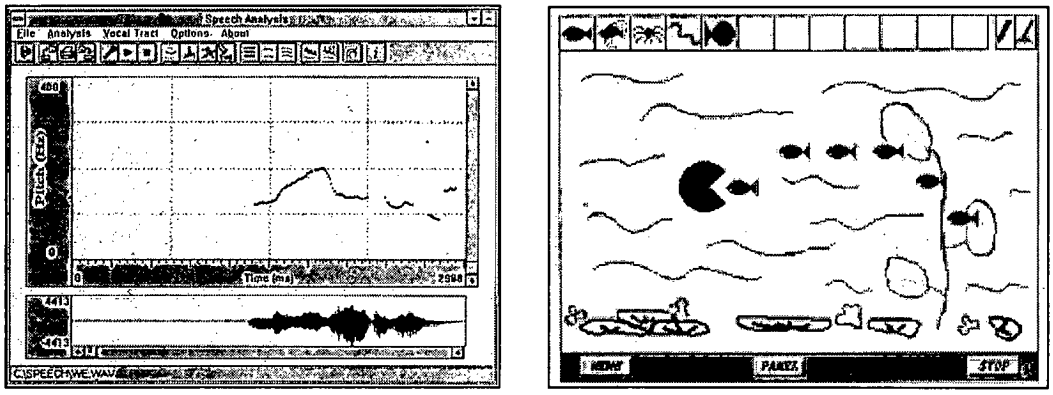


Figure 3.6. Two examples of frequency/time planes (left: Dr. Speech Science; right: Visual Speech Apparatus)

When a trail of the F_0 track is left on the screen, like in the first example in Figure 3.6, an F_0 contour is shown, giving useful information for improving the control of prosody.

Also, "pitch games" are often realised by using a frequency/time plane, as in the second example of Figure 3.6 above, where the task is to collect "targets" on the screen. For users that tend to have a monotonous voice it may be appropriate to position the targets at the extreme of the screen, to encourage variation, while for users with random pitch movements it may be appropriate to align the patterns on a horizontal line.

Sometimes F_0 is represented by an object moving horizontally, for example along a piano keyboard (such as in the IBM SpeechViewer 2). The placement of the chosen object in most cases is linearly related to the fundamental frequency, although some systems display logarithmically, like for example in the case of the piano keyboard.

Notice that the term "pitch", as used by many speech rehabilitation system manufacturers has to be interpreted as "fundamental frequency". All systems simply measure the fundamental frequency and display it regardless of fact that the perceived pitch depends on other factors as well, as discussed later in Chapter 6.

Vowels and diphthongs

Vowels are represented in many ways: by bar spectrograms or grey-scale spectrograms, as in Figure 3.7,

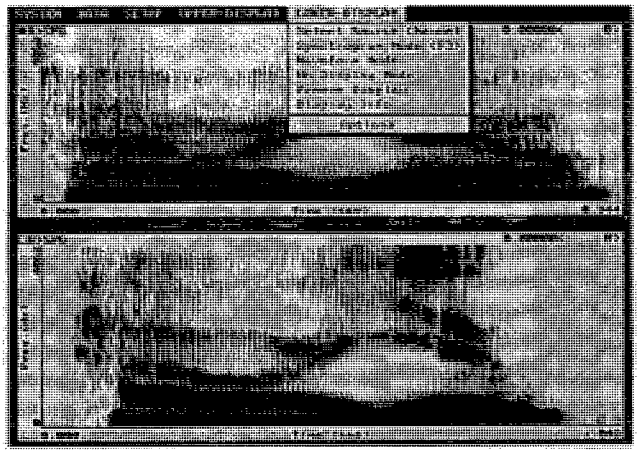


Figure 3.7. Grey scale spectrogram (Kaway ProTS)

or by variations, such as in the case of the "spectral movie" (Nickerson & Stevens, 1973) where the spectral distribution is rotated, reflected and enclosed, (see Figure 3.8), to form an "object" which changes shape with the speaker's voice.

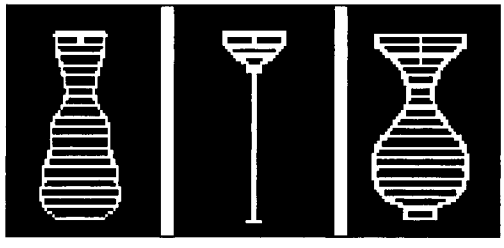


Figure 3.8. Spectral movie[1]

[1] Images from the original article. A similar display is today used in the Panasonic VH-9500.

Other representations are bi-dimensional F1-F2 planes, as in Figure 3.9,

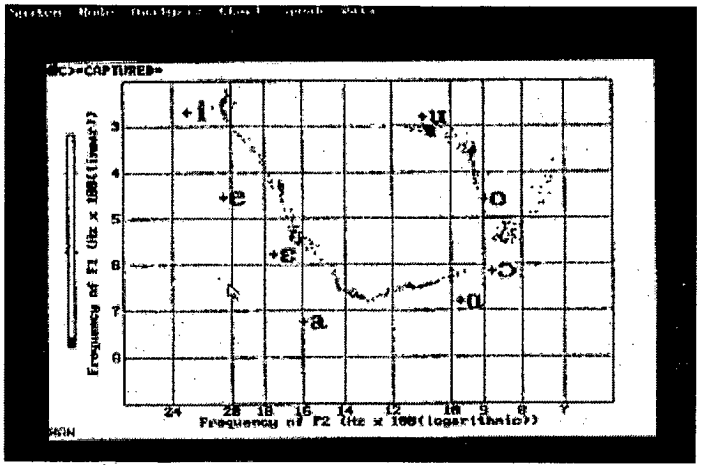


Figure 3.9. F1-F2 plane (Kay Sona-Match 4327)

or more recently by showing a real-time animation of the vocal tract section (for example Rossiter, 1994) or non-real-time as in Figure 3.10.

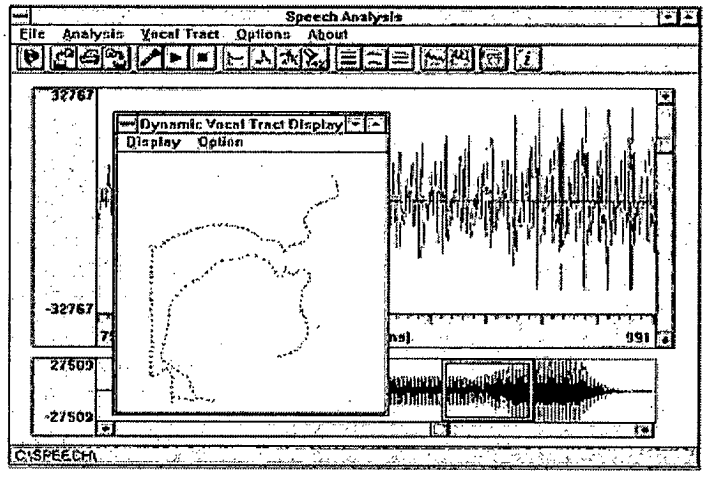


Figure 3.10. Vocal tract section (Dr. Speech Science)

Other displays use a categorical (versus continuous) representation of vowels, for example by illuminating the vowel being pronounced, as in Figure 3.11,

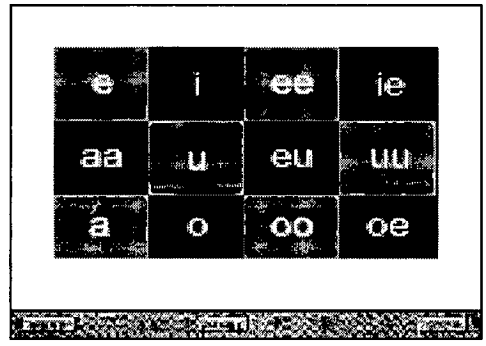


Figure 3.11. Categorical display of vowels (Visual Speech Apparatus)

or selecting four vowels and using them for moving up, down, left or right inside a maze, as in Figure 3.12.

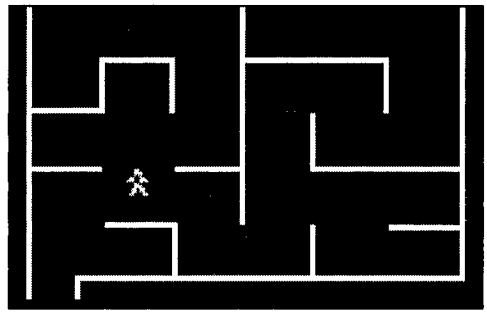


Figure 3.12. Vowel maze (IBM SpeechViewer I)

Diphthongs are represented in some system (such as the IBM SpeechViewer 2) by simply showing the variations of formants in a grey scale spectrogram like the one in Figure 3.7.

Consonants

Generic fricatives may be shown using a "vocalisation" display as described before in this section. Selected consonants may be shown categorically using the same kind of feedback used for vowels. Voicing errors, omission errors and phoneme order errors can be detected by a "phoneme chaining" game feedback like the one in Figure 3.13, where the penguin jumps to the next step only if the correct phoneme of the chain is pronounced. The chain of phonemes may contain vowels as well as consonants, in order to form real words if desired. However there is the problem that in case of error no indication is given about its type and amount.

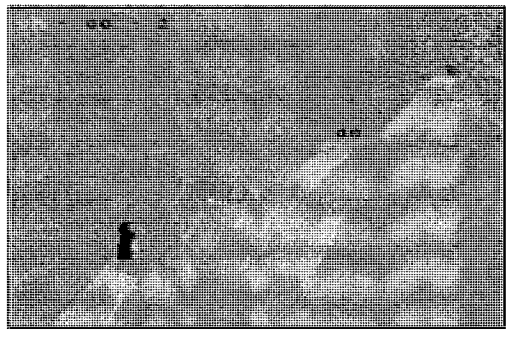


Figure 3.13. "Phoneme chaining" game (IBM SpeechViewer 2)

To deal with place of articulation errors such as the production of /ʃ/ (like in 'shoe') instead of /s/ (like in 'sea'), or the reduced distinction between them, "S indicators" are often used, where a light signals the /s/, whose spectra contains more high frequency components than /ʃ/.

Some systems, in addition or in substitution of a feedback on consonants, give some examples, generally in the form of printed material or sometimes, in expensive systems, as moving images on videodisks, as shown in Figure 3.14.

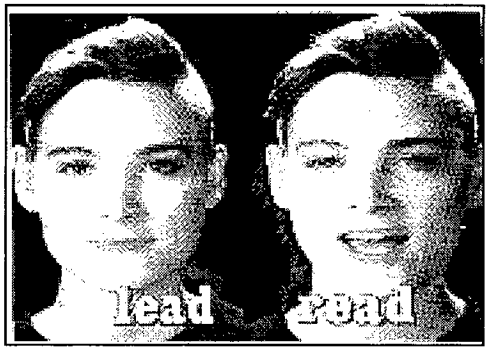


Figure 3.14. Moving pictures with examples of consonants (Kawai ProTS laser videodisc).

Voice quality

Nasalization is shown only in systems having the appropriate sensor (accelerometer) and is represented by a light indication, or a picture of a face with the nose changing colour or size. On Phonation, some systems give the amount of jitter and shimmer (which can relate to harsh or creaky voice) on bar meters. Notice that voice quality may affect the perception of pitch, but no system measures this characteristic of fundamental frequency variation.

Timing

As discussed in section 2.4.6, hearing-impaired people tend to speak much more slowly than normal-hearing people for two reasons: longer segments, especially vowels, and inappropriate use of pausing. It appears that with some published systems one possibility is to show these aspects by the use of a "vocalisation display". Nevertheless for the user it may be a problem to interpret the data in a meaningful way, since there is the complication that often deaf speakers shorten some syllables by omitting some segment, although the total time of an emission is longer than it should be. It appears that without segmental analysis it may be difficult to achieve a reasonable feedback.

Pausing

Pausing can be displayed on a "vocalisation" display, even if it may be difficult for the users to understand how to change their speech in order to reduce this problem.

Breath control

Practising with sustained sounds is one method for obtaining better breath control. A "vocalization display", a spectrogram display, or even a pitch track are used in this case to show how the phonation is maintained for the requested amount of time.

Rhythm

The perception of rhythm depends on many aspects of speech production, such as loudness, pitch, vowel quality and duration of segments.

Speech Visualisers

Some displays show many speech features at the same time, often presented as a combination of a speech spectrogram and a pitch tracker. An interesting variation of a speech spectrogram was developed by Watanabe (1985, 1995), where connected speech is converted into colour pictures. In this system the lowest three formant frequencies are extracted from voiced speech, and are converted to three primary colour signals. In unvoiced portions, colourless and dapple patterns are displayed. The author claims that the reproduced pattern is not only beautiful, but also easy to understand intuitively.

3.4 Conclusion

This chapter has reviewed the most important studies on visual feedback for hearing-impaired speech rehabilitation, and gave examples of feedback from many published systems. It appears that, as a general rule, visual feedback has been designed using a practical and experimental approach, instead of being the result of some user-oriented research. Although the feedback included in these systems is currently used by a large number of therapists to help their clients to enhance their speech, this does not mean that the feedback being used is the most appropriate.

In the next chapter the problems that therapists and hearing-impaired users encounter while using these systems are discussed, and a new approach in the design of visual feedback is proposed.

CHAPTER 4

A Novel Approach to Designing Visual Feedback for the Rehabilitation of Hearing-Impaired Speech

4.1 Introduction

The previous Chapter has described previously published studies in the field of visual feedback for voice rehabilitation, showing examples systems which have resulted from these studies. Systems for voice rehabilitation are used every day in many speech therapy laboratories, but users and therapists often encounter problems that have still to be solved. These problems are discussed in Section 4.2.

A new approach for designing visual feedback for voice rehabilitation of hearing-impaired speech is therefore proposed in Section 4.3. In this approach, an intuitive correlation between different visual dimensions and different features of speech is investigated by means of experiments, using as background extensive research results from the field of visual interface design. The results of the

experiments are then used to design a set of feedback methods for speech rehabilitation purposes, taking into account comments from therapists familiar with actual speech rehabilitation systems.

Furthermore, new graphic techniques such as multimedia and virtual reality are now available at a reasonable cost on today's personal computers, and these techniques may greatly increase the effectiveness and enjoyment of visual feedback. No research has been done to date to test their effectiveness in the field of speech rehabilitation. Visual interface design techniques are discussed in Section 4.4. The various modalities used to display information on a screen are reviewed, and guidelines on their use are reported. Section 4.5 compares the guidelines on visual interface design with the studies on visual feedback for voice rehabilitation, and the actual implementation in systems currently in use. Section 4.6 highlights the need for experiments, since the problem of finding intuitive and effective visual feedback remains unsolved and traditional visual interface techniques can probably be exploited in more effective ways. Furthermore, novel techniques such as virtual reality has not previously been tried before in this field, are therefore candidates for designing effective and motivating visual feedback. The actual design, implementation and results of these experiments are reported in Chapter 5.

4.2 Problems with Visual Feedback currently used for Hearing-Impaired Speakers

Speech rehabilitation systems using visual feedback have been used for many years by speech therapists with their clients. These systems (Nickerson, 1973; Brooks et al., 1981; De Bot, 1983; Arends, 1993; Pratt, 1993) offer help in the speech rehabilitation process. However there remain problems to be solved. Some of the problems listed below were underlined in a survey of twelve British therapists who agreed to record a selection of specialist views on visual feedback technology for the hearing impaired. These specialists were contacted and interviewed in "special interest groups" held in speech therapy laboratories and schools for the deaf. Specialists included speech therapists for the hearing impaired, teachers of the deaf, linguist and student speech therapists. Of the twelve specialists who contributed to the survey (see "Survey questionnaire" in Appendix A), six were familiar with the SpeechViewer (IBM), three of these had also used a Visispeech (Mill Grant Wells) and one was also familiar with Cspeech (Univ. of Cambridge) and Micronose (Medical Physiscs Dept., Wakefield). Only one of the participants (a student) had not come across any such systems. The problems highlighted cover both aspects of visual feedback to the user, and other issues such as courseware and cost.

Negative reinforcement

One of the most significant problems identified by therapists is the tendency for systems to negatively reinforce some aspect of speech production and disallow good speech. This is caused mostly by two reasons. One of them is the incentive to strain the voice, caused by graphic feedback where more attractive graphics are shown at the extremes of the voice range (for example, a balloon expands in proportion with the voice intensity, and if the voice is too loud, the balloon bursts). This leads the user, especially children, "to tend to just voice indiscriminately to make the pictures activate" in order to enjoy the amusing effect, with possible injury to the voice. The other reason is inaccuracy in the speech analysis, as explained later in this section.

Lack of motivation

Most of the interviewed therapists expressed a preference for a wider variety in displays and games, both because children become bored easily unless their attention is gained with "surprises", and also because a wider variety of displays can be used as a basis to create new situations and games between the therapist and the child. Some therapists say that the system they use "gives you a choice of animation / picture you want, but it only gives you a choice of pictures and not a choice of games". One therapist commented they would like to see "multimedia type" images to explain the lesson in a more enjoyable way to increase motivation. Another therapist would like to have more "visual material". Another commented: "There is the need of moving out of the 'computer technical' field and move into the 'arcade game' type of visual reinforcement to users".

Frustration

The survey also showed that it is necessary to avoid provoking frustration in the user. Systems should avoid confronting the user with targets representing normal speech, which are often too difficult to "hit", and for this reason became a source of frustration. Feedback must be more positive, e.g. reinforce and reward small steps in the right direction rather than give negative response to anything that is not 'correct' (use for example of relative responses like: 'Better than previous attempt'; avoid negative responses such as: "Incorrect!"). "Undermining of client's confidence and negative feedback lead to a great tension and slower self-esteem". It is more appropriate to adapt the scope of the session in a way that the user can reach the target and increase the difficulty gradually, session after session. Furthermore the system must be able to deviate from the norm, for example, if a child cannot pronounce /s/ and is trying to say "seven" and can only say "even", when the child progresses to "theven" the computer must give praise. Therapists were unsure about the usefulness of a "score" to show how well a child has done, since it "emphasises qualitative feedback". "Graded feedback would be useful, but a 'score' tends to imply that this 'normal' is eventually acceptable". Many profoundly

deaf people will never achieve normal targets". One therapist takes the view that scoring may be a good idea, provided that the scores are always high, "or they'll be put off'. Another therapist thinks scoring is not useful because "if the client is self-monitoring they generally have an idea of how successful they are". Note that the problems with scoring contrasts with the proposal mentioned in Section 3.2.1, where a "video-game type" of visual feedback, including scoring, is considered appropriate for motivating users.

Rewards should also be designed with care. "Something interesting may happen when they get it right, possibly for young children, but this implies an 'absolute rightness' whereas I would prefer the system to reflect degrees of progress". A reward may be useful "but be careful that this wouldn't discourage them in any way when they get it wrong, i.e. within reason".

It has been shown that mood plays an important role in learning. Not only do chronically depressed patients have difficulties in learning (Beck, 1967), but to learn in an experimentally induced state of depression has been found to show an efficiency decrease of 30% in comparison to a neutral or good mood status (Ellis, Thomas & Rodriguez, 1984; Leight & Ellis, 1981; figures based on studies with normal-hearing subjects).

Inflexibility

Some therapists complained that the system they used was not flexible. One said that they prefer "home-made" materials, and the system should allow therapists to tailor the feedback as they think appropriate for their clients. Therapists would like to be able to choose the level of detail in displays such as stress / intonation. They would all like to see detailed data about clients' speech for their own diagnosis purposes (such as spectrograms, fundamental frequency, vowel formant frequencies). Some therapist complained about the limitations in scope of some system (C-speech), the unsuitability of some other system for children (VisiSpeech) and the fact that in another system (SpeechViewer) was necessary to set the sensitivity differently for each user. Some therapist think that courseware is more useful where a set of flexible modules exists, dealing the main pronunciation problems, each independent by the others, instead of having a set of modules to be taken in a particular sequence, forming a structured "curriculum" or teaching programme. "Too rigidly structured lessons could cause problems, less progress, less enjoyment". Furthermore, speakers should be allowed to pick and choose the lesson they want to do each day. All therapist would like a larger variety of programmes, for example a wide choice of intonation patterns, stress patterns, full range of vowels.

Difficulty to understand how to change the pronunciation to get it right.

Some feedback does not provide information about how to correct the pronunciation. This problem is often more evident with categorical displays. In case of vowel quality, some therapists proposed an indication of the quality of articulation would be needed to bring the vowel closer to the target, such as tongue position and general articulatory position. Some users find the goal of the drill "difficult", in the sense that it is not clear what they have to do.

Ease in navigating

If a system is easy to use and navigate in different screens, the user would "press the buttons" for controlling the computer, increasing motivation. Generally this is not the case.

Not fast enough

Some therapist would like a shorter analysis delay time, "immediate", "the quickest the better". However this applies only in the cases where a "delayed-feedback" is not required.

Problems with multidimensional displays

Some therapists think it is not a good idea to make talkers learning one feature (e.g. intonation) become aware of mistakes in other aspects of pronunciation (e.g. segmental), because this may be confusing and de-motivating. "Learning one aspect of language at the time is usually more beneficial than mixing objectives. Perhaps at an advanced level a user may cope". "Displays should be not too intricate, with clear easy recognisable graphics, aesthetically pleasing to motivate". The systems with which therapists were familiar generally did not generally made use of multidimensional displays, with the result that the generally negative response on multidimensional displays is in contrast with some of the guidelines found in the literature, as shown in Section 3.2.3. An example of a system using multimensional displays is the Speech Spectrograph Display (Maki, 1983). For more information on the topic see Powel & Maassen (1987), and Watanabe, Ueda & Shigenaga (1985).

Categorical displays

Most therapists think that categorical displays should be avoided because they are frustrating and they do not give indications of how to correct the pronunciation. In the case of vowel displays, targets should be faded instead of being too definite.

Inaccuracy

Some therapist complained about the accuracy of the systems they were using (C-speech). One therapist complained about having to satisfy the system's demands for right & wrong "when to my own ears the productions were perfectly OK". The accuracy of the system is also very important if the system has to be used autonomously by the user without the therapist always present. Some therapists complained about the accuracy of a fricative detector (SpeechViewer) that declares some sibilant fricatives to be voiced.

High cost

Most therapists complained about the high cost of many systems (IBM), permitting them to buy fewer systems than they needed.

Better if portable

Some therapist would like the system to be portable, in order to take it into classrooms and homes.

Discussion

Attempting to categorise the reasons for the problems listed in the previous section, four areas are identified:

- Evaluation / feedback from users (deaf related issues)
- Human factor / system design
- Research in visual / auditory perception
- Speed / accuracy of speech processing

Table 4.1 lists the problems seen above and relates them to the areas involved with these problems. The column on the left lists the problems, while the row at the top lists the areas.

	Evaluation / feedback from users	Human factor / system design	Research in visual / auditory perception	Speed / accuracy of speech processing
Negative reinforcement	♦			♦
Lack of motivation	♦	♦		
Frustration	♦			♦
Inflexibility	♦			
How to get it right?	♦		♦	
Difficult navigation		♦		
Not fast enough				♦
Multidimensional displays	♦	♦	♦	
Categorical displays	♦			
Inaccuracy				♦
High cost				♦
Portability				♦

Table 4.1. Problems and related areas

Evaluation / feedback from users (deaf related issues)

Problems such as negative reinforcement, lack of motivation, frustration, inflexibility, difficulty to understand how to get it right, and visual display aspects such as use of categorical and multidimensional displays, may depend on more than one reason, but the cause is probably an incomplete evaluation of system effectiveness, and limited use of feedback and advice from therapists and users. From Watson (1991): "The reason for the failure of many of the developed aids not reaching the market can be attributed to many things. But perhaps the most important are the too high expectations of the engineers and therapists on the usefulness of the aid and the lack of rigorous evaluation of the aid in therapy clinics".

Human factor / system design

Human factors and visual system design helps in building usable, motivating systems. The extensive research (see example Marcus, 1995) carried out in this field is an important source of hints and guidelines that needs more consideration in designing visual displays for speech rehabilitation.

Research in visual / auditory perception

Little research has been reported on the differences and links between visual and auditory perception in deaf people. The problem of how to visualise speech is still unresolved. If a display does not give a clear and intuitive indication about how to change the articulation to get it right, this means that probably that display is not optimal.

Speed/accuracy of speech processing

Speed and accuracy of speech processing are generally linked. Accurate and fast processing is expensive. Inaccurate speech analysis negatively reinforces speech and slow response is the cause of boredom and lack of motivation. Most systems require the addition of plug-in accelerator boards. This requirement is not only expensive, but also limits the possibility of building a portable system, since most portable computer systems (such as laptops) cannot fit a plug-in speech processing accelerator board.

Conclusion

Bernstein (1989) believed that work to provide adequate speech training for hearing-impaired individuals is still in the growth stage, that there is room for great improvement. It looks like today the situation is not changed dramatically. It seems a valid consideration by Watson & Kewley (1989)

who, in a discussion on training aids for the deaf, state that the large variety of displays in use (from single points shown on a vowel space, to bar graphs, to animated cartoons, and even interactive video-games), "were determined more on the imagination and creative skills of the authors than on the particular aspect of speech to be trained - or on any hard evidence favouring one or another type of display". In fact, published studies are rather limited in number, and sometimes appear as not too rigorous, especially if compared with the quantity and often strictly accurate research on perception of normal-hearing people. Some inaccuracy (or lack of adequate information) was noticed when comparing studies on displays for the deaf with studies in experimental psychology (for example Treisman, 1986), from which it appears not the case that most visual features of an object are coded in early vision (see later in Section 4.4.9), as stated in the literature (Arends, 1993). Furthermore, properties such as colour and brightness, difficult to perceive separately, are catalogued in the literature as unique dimensions, usable in multidimensional displays (see Section 3.2.3). It should be said that these inaccuracies have not led to harmful results, since it appears that in the end the majority of graphic display designs for the deaf do not take into account these limited theoretical studies at all. The Visual Speech Apparatus (Arends, 1993) is a significant example. It is a well known system, based on recommendations by Povel and Maassen who, considering the limited success of the numerous existing visual aids, as reported by Lippman (1982), proposed a practical approach as the best one for building an "effective visual aid" " (Povel & Maassen, 1987). The motivation was, about the ideal speech visualiser "...even a superficial study of the differences between visual and auditory perception on the one hand and of the coding of speech in the acoustic signal on the other, makes clear that this is a most complicated undertaking." It is interesting to notice that the same issue is today still unresolved in the wider field of Human Computer Interaction. From the February 1994 workshop on HCI, page 19: "...fundamental theory on Human Computer Interface is scarce. Some attendees felt that HCI is an applied field and does not need its own theory. [...] Therefore it is proposed that significant progress in HCI can be achieved by engaging in fundamental research aimed at understanding more deeply the basic limitations of human information processing [...], and seeking to solve the open question of whether or not there is any theory, or set of theories, applicable to the design and development of interactive systems. A few classic psychological or psychophysical studies do exist, such as Miller (1956) and Fitts (1954), and others have wondered about the prospects for such an approach (Newell & Card, 1985). Even so, available theories are not sufficient to fully understand and predict human interaction phenomena, especially with respect to computer use. More work is needed at the boundary between computer science and cognitive science to settle the question of the existence of applicable theories for HCI, and their relative importance."

Nevertheless, extensive studies have been carried out in the field of visual interface techniques, and the results of these studies have been taken into account in the design of a large number of systems, leading to an effective enhancement of these ones. This vast experience can be used to advantage as

the starting point for the design of visual interfaces for the hearing impaired, being this specific area one of the many possible applications using visual interface techniques. Merging this experience with feedback from therapists and users of the actual visual speech rehabilitation systems, integrating the results of novel experiments about links between visual stimuli and speech features, and implementing some effective and efficient speech processing method, may solve some of the problems seen above and help to enhance systems using visual feedback for deaf speech rehabilitation.

4.3 A New Approach

As discussed in the previous section, the design of many systems for hearing-impaired speech rehabilitation is based on a practical approach. The reason given is that a more theoretical approach was too difficult to follow, due to the lack of knowledge in the field of the association between visual and auditory perception.

As also seen in the previous section, it appears that feedback from therapists and users, was not taken into account as much as it deserved. Similarly, guidelines about visual interface design were sometimes ignored, resulting in user interfaces that were not easy to use, or not as effective as they should be. This approach has resulted in systems that exhibit a series of problems. A new approach for designing visual feedback for voice rehabilitation of hearing-impaired speech is therefore proposed, as shown in Figure 4.1.

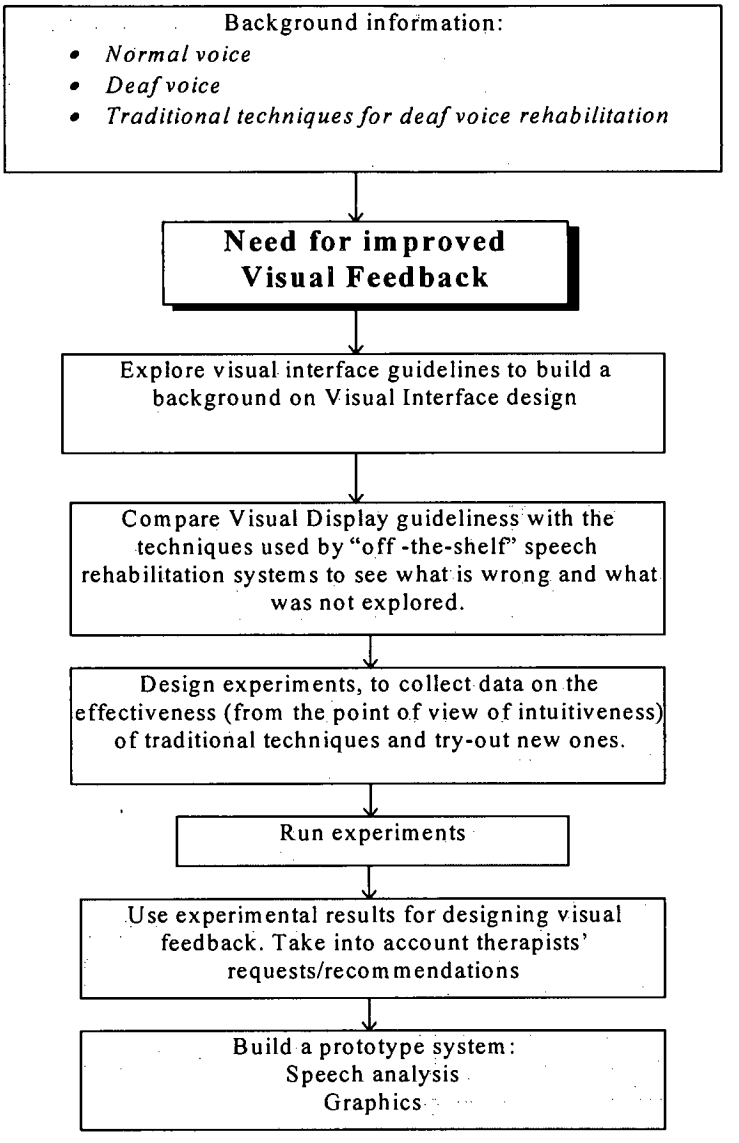


Figure 4.1. A new approach for designing visual feedback for voice rehabilitation of hearing-impaired speech

The approach is based on background knowledge on normal voice, deaf voice, and traditional techniques for deaf voice rehabilitation, and follows these steps:

1. Explore visual interface guidelines to build a background on visual interface design: familiarise with the techniques for displaying information, such as positioning, grouping, highlighting, using colours, and with the guidelines about the use of different modalities, such as text, graphics, images, animation, video, virtual reality.

2. Compare the guidelines about visual interfaces with the visual feedback methods for the published hearing-impaired speech rehabilitation systems. Find out what visual feedback used in published systems work in a satisfactory way, and which do not. Highlight inconsistencies, and methods that were not explored.
3. Merge this information by selecting a number of modalities to display information that may be appropriate for use as visual feedback for hearing-impaired speech rehabilitation.
4. Test the response of hearing-impaired speakers to these modalities with a set of experiments, in order to establish the most intuitive match between different visual stimuli and different speech features.
5. Compile a list of results to "score" the intuitiveness and effectiveness of each visual stimuli as a method for representing the different speech features. Take into account therapist's requests and recommendations.
6. Select one or more of the most promising pairs "visual stimuli - speech feature", giving priority to those that may give an effective improvement to the methods used traditionally, and build a prototype system, also considering real-time response, accuracy of speech analysis and final cost of the system.

The next section deals with the first of the steps explained above, in order to provide enough background information on visual interface techniques.

4.4 Introduction to Visual Interface

Recent reports estimate that more than half of the cost of a new computer system is attributed to the user interface (Bass, 1993; Myers & Rosson, 1992). The importance of the interface was also emphasised by a industry representative research group, who concluded that "if the interface is ineffective, the system's functionality and usefulness are limited; users become confused, frustrated, and annoyed; developers lose credibility; and the organisation is saddled with high support costs and low productivity" (Nolan & Norton, 1992).

A large part of the user interface consists of the *visual interface.*

This section describes visual interface in general. The techniques and the guidelines about displaying information on a screen are therefore reviewed, with the goal of building a theoretical basis of this matter, to be used as a starting point for the specific application being dealt with.

The design of a visual interface in a computer system probably has received more attention in the guidelines of human-computer interaction than any other aspect of the interface. There are many examples that show the importance of the design of a visual interface. Tullis (1988) has verified that redesigning a display used to test telephone lines gave a reduction of 40% in the time necessary for interpreting the data shown in it. This redesigned test resulted in a reduction in work equivalent to 79 people a year for the telephone company that adopted it. In a similar way Keister and Gallaway (1983) found that redesigning a set of displays resulted in a reduction of 25% in the time required to process the data, and a reduction of 25% in error rates. In a study on more than 500 displays used in travel agencies, Tullis (1988) calculated that the time spent in interpreting the data using the display that was judged the worst one was 128% longer than the time spent using the best one.

These studies concerned visual interfaces based on alphanumeric displays, only sometimes enhanced by some sort of simple graphic (such as frames to group textual data), also known as "semi-graphics", and with rather low resolution.

In recent years, Graphical User Interfaces (GUI) have been developed which allow users to deal with computers using simple actions, like pointing icons with a hand-held device known as the *mouse.* Millions of people who were previously novices with computers, are now able to control applications like spreadsheets or word processors. This is due to the enhanced graphical possibilities of personal computers that make it possible to simplify the use of applications through the use of transparent metaphors. For example, in the display of a word processor programme objects often appear whose function is clear, without the need of explanations: clicking on the symbol of the printer produces a

hard-copy printout of the text, clicking on the image of a pair of scissors the "cut" function of "cut and paste" is performed, and so on.

In comparison with the displays which make use of alphanumeric text only, graphical displays enhance the possibility of presenting information in clearer and more intuitive ways. There remains the problem of how to place information, how to codify it (the way to present it), and how to balance it in an effective way.

Many articles present guidelines on screen design. Important studies are the ones by Galitz (1989), one of the most comprehensive about alphanumeric displays, while in the field of graphical displays the work of Smith & Mosier (1986) and Marcus (1995) are often used as reference books. (Sources: Tullis, 1988, Ui-Design, 1996, and others). About alphanumeric displays Cropper & Evans (1968) discuss colour, grouping, consistency, and density; Engel & Granada (1975) give about 60 guidelines addressing display formats (e.g. highlighting, layout) and content (e.g. feedback, labels, messages); Pew & Rollins (1975) give about 35 guidelines developed for the US Dept of Agriculture addressing standard format, menus, data entry forms, output data, and error messages; Stewart (1976) discusses six screen design factors (logical sequencing, spaciousness, relevance, consistency, grouping, and simplicity) and 6 colour techniques (alphanumeric, colour, brightness, spatial, shape, and flashing); the MIL-STD-1472C (1981) gives about 37 guidelines addressing standard formats, grouping, updating, coding, density, and data presentation; Williges (1981) compiled about 110 guidelines from 13 other documents addressing information coding (colour, shape, brightness), density, labels, format (prompts, tables, graphics, text), and layout; Bailey (1982) discusses grouping techniques, standardisation, messages, format, labelling, multi-screen displays, presentation of coded information, screen organisation, feedback, highlighting, and cursor design; Pakin & Wray (1982) discuss of format elements, colour, content, layout, style, and word use; Lockheed (1983) contains about 72 guidelines or standards addressing display areas, conventions, alphabetics, numerics, layout, lists, abbreviations, labels, terminology, and use of colour; Tullis (1983) reviews literature on four basic characteristics of alphanumeric displays; overall density, local density, grouping, and layout complexity, and presents measures for each and applies them; Burroughs (1986) contains about 60 guidelines or standards developed for InterPro software series addressing a standard layout, terminology, letter case, justification, headings, menus, field alignment, check-off lists, and field labelling; Galitz (1989) gives several hundred guidelines addressing general screen design (e.g. placement, fonts, messages), data entry screens (e.g. headings, alignment), inquiry screens (e.g. organisation, justification), menus and colour.

Danchak (1976) discusses graphical screen designs for power plants, including coding (numeric, textual, shape, colour, blinking), density, word length, grouping and preferred quadrant; Peterson (1985) discusses menus, prompts, information presentation, data formats, text, and messages; Smith

& Mosier (1986) give about 300 guidelines addressing data displays, including text, data forms, tables, graphics (scaling, scatterplots, line graphs, bar graphs, pie charts, pictures, maps), format, coding (highlighting, colour) and display dynamics (selection, framing, update); NeXT (1992) contains the guidelines on the NeXTSTEP user interface; Apple (1993) gives a description of the standard MacIntosh user interface, including icons, palettes, pointers, selected objects, windows, scrolling, pull-down menus, dialogue boxes, buttons, check boxes and radio buttons, dials and alerts; IBM (1993) contains guidelines on object-oriented interface design; OPEN (1993) contains the Style Guide of OSF/Motif; Marcus (1995) discusses format and grids, typefaces, colour, screen metaphors, sign-on, windows, menus, icons and cursors, animation, sound, across Windows, NeXTStep, Mac; Microsoft (1995) contains the guidelines on the Windows interface; Mullet & Sano (1995) state six major principles: elegance and simplicity, scale/contrast/proportion, organisation, module, image and style.

The abundance in guidelines regarding screen design do not correspond to an abundance of empirical evidence about screen design. Many screen design issues have to be addressed empirically, especially those that make use of bit-mapped graphics. Table 4.2 summarises a selection of empirical studies addressing screen design (Source: Tullis 1988).

Reference	Types of Screens	Key Findings
Benbasat, Dexter & Todd (1987)	Tabular (A/N) and graphical presentations of marketing data in colour or monochrome; measured decision-making performance.	Graphics are more useful when evaluating data to detect promising directions; tabular displays are more useful in tasks requiring precise computations. Colour resulted in fewer iterations to complete the task.
Burns, Warren & Rudsill (1987)	Current and reformatted A/N screens for the Space Shuttle; changes included grouping of related data, use of indentation to indicate subordinate relationships, and consistent use of abbreviations.	Reformatted screens resulted in 30% reduction in search time for non-experts and no change in search time for experts; both groups showed greater accuracy with the reformatted screens
Callan, Curran & Lane (1977)	Six A/N formats for presenting Navy tactical information; varied the number of display items from 6 to 40.	Search time increased linearly with the number of items.
Card (1982)	A/N menus with different semantic bases for groups: alphabetical, functional, or random.	Time to select an item from the menu was fastest for the alphabetical grouping and slowest for the random grouping.
Dodson & Shields (1978)	Three A/N formats for Spacelab displays: 30%, 50% and 70% overall density.	Search time increased significantly as overall density increased.
Duchnicky & Kolers (1983)	Scrolling text displays that varied in character density (40 or 80 per line), line length, and height of scrolling window; measured reading rate and comprehension	Full-width and two-thirds-width screens were read faster than one-third width. Character density of 80 per line was read faster than 40. Taller windows yielded slightly faster reading.
Keister & Gallaway (1983)	Original and redesigned A/N screens for an NCR software package; redesigns involved changes in screen format and content, consistency, data entry procedures, error messages, and on-line help.	Redesigned screens resulted in 25% less time to complete transactions, 30% less data entry time, 25% reduction in error rate, and 32% reduction in time spent correcting errors.
Kolers, Duchnicky & Ferguson (1981)	Single and double-spaced displays of text.	Single spacing required more eye fixations per line, resulted in fewer words read per fixation, and required longer total reading time.

LaLomia & Coovert (1987)	Line graph or tabular displays of data from standardised test questions. Used four tasks: locating a data value, trend analysis, interpolation, and forecasting.	Locating a data value was faster with tabular displays. Trend analysis and forecasting were faster with line graphs.
Ringel & Hammer (1964)	Nine A/N formats for tables presenting the status of battlefield units; varied the number of lines of data and the amount of space between the lines.	Search time increased with the number of lines of data and decreased slightly with less space between the lines.
Schwartz (1986)	Six A/N formats for presenting the status of systems within a hypothetical power plant; varied local density, grouping, and item alignment; subjects scanned each display and made decisions based on five critical data items.	Size of groups of characters and alignment of data items were the best predictors of scanning time; number and size of groups were the best predictors of perceived ease of use.
Schwartz & Howell (1984)	Graphic and numeric representations of hurricane tracking data.	Under time pressure, subjects reached a better and faster decision using graphic rather than numeric data; under self-pacing there was no difference.
Streveler & Wasserman (1984)	Seven pairs of A/N formats for presenting a variety of data; studied effects of grouping, alignment of data, position on screen, and ordering of lists.	Search time decreased significantly with clearly defined groups and well-aligned items. Shortest search times for targets were found in upper-left quadrant and longest times in the lower-right.
Tullis (1983)	Four designs for results of tests on telephone lines: "narrative" A/N, "structured" A/N (with better-defined groups of data), monochrome graphic (with a simple schematic), and colour graphic.	After practice, time to answer questions about the display was 40% shorter for the "structured" A/N and both graphic formats than for the "narrative" A/N format. Subjects expressed preference for the colour graphic format.
Tullis (1984)	52 A/N formats for presenting airline and lodging information and 15 A/N formats for presenting information about books; measured 6 characteristics of the formats: overall density, number of groups, size of groups, number of items, and item alignment.	Multiple regressions using the format variables as predictors accounted for 51% of the variance in search time and 80% of the variance in rated ease of use; best predictors of search time were number and size of groups, and of rated ease of use were local density and item alignment.
Wolf (1986)	Five A/N formats for presenting lists of the names of host computers available in a network; manipulated order of names (alphabetical by row or column) and grouping.	Shortest search times and best ratings were for a format with clearly defined columns of names, alphabetically ordered within each column.
Yorchak, Allison & Dodd (1984)	Fixed 2-dimensional graphical display and simulated user-defined 3-dimensional display of satellite altitude and coverage area.	Subjects preferred the 3-D displays, although there were no differences in speed or accuracy of display interpretation.

Table 4.2. Selected empirical studies of screen design

The following sections describe the guidelines resulting from the studies reported above.

4.4.1 Amount of information presented

Almost all of the guidelines specify that the amount of information to be presented to the user has to be minimised, and only that information necessary should be shown (Smith & Mosier, 1986; Galitz, 1989; Marcus 1995). Other authors went into this idea in depth, specifying *density* values. Danchak (1976) proposed not to take more than 25% of the usable area in the display. In one of his studies it is reported that displays judged qualitatively "good" had an average use of the usable area of 15%. On the other hand, in a series of guidelines for the design of displays in the Spacelab, NASA (1980) affirms that density should not generally take more than 60% of the available space. Empirical evidence on this point is generally consistent: provided that the information necessary for executing a due task is present, the human performance tends to deteriorate with the increase of display density (Callan et al., 1977; Dodson & Shields, 1978; Ringel & Hammer, 1964). Even if the effectiveness of a particular display may be improved by formatting information in a different way, it remains a fact that

the optimum amount of information to show is that necessary for executing the requested goal - no less, no more. Which information is necessary to execute the requested goal can only be determined by means of a complete analysis of the goal that is expected from the system user (Tullis, 1988). Once the relevant information to display has been determined, different techniques can be used for placing it adequately, so as to avoid overcrowding the screen, as discussed in the next section.

4.4.2 Placement of information

Placement on the screen should be done in such a way as to give the perception of as much information as possible (taking into account the considerations given in the previous section) without the need to move the observation point. To clarify this idea it is useful to introduce the concept of *viewing field.* It is possible to distinguish three viewing fields:

- *stationary field* (the field within 30 degrees of the *viewing angle*) where it is not necessary to move the *eye fixation point* to perceive the status of a certain pieces of information;
- the *eye field* (the field within 30 and 80 degrees), where movement of the eyes is necessary; normally it is advised to place one of the displays in the area closer to the stationary field (*direct vision field*), and the others in the more peripheral areas; displays located in the peripheral areas require attention through *pre-attentive* processes (Neisser, 1966); these processes imply that the eyes have an automatic reflex caused by the perception of a movement. Therefore eye movement is not under the user's control, because this happens before the subject is aware of it;
- the *head field*, where a movement of the head is necessary. In this case display is outside the peripheral field, and the pre-attentive processes with the relative automatic reflexes are not enabled.

To take advantage of the visual field, knowledge of the importance of information in each display is essential. Information used more frequently should be placed in more accessible areas, while that used more rarely can be placed in less handy areas. *Task analysis* is useful to decide the importance of the various displays according to their function and frequency of use: this method helps in categorising display in three groups, primary, secondary and emergency. These three categories should have a different position in the visual field. Sanders (1977) proposes to place primary displays in the stationary field, and the secondary and emergency ones in the eye field. Displays belonging to the same function should also be grouped together, and not mixed with those relative to other functions. Displays used for emergency purposes should be isolated from the others, and their information should be easily identifiable. As a general rule, displays should be placed in such a way as to make it easy for the user to comprehend their functions, both singularly and in relation with other displays. This may be accomplished by placing the display according to a geometrical structure that resembles the physical position of the functions that are monitored. The *Gestalt theory* on information grouping

studies these problems. According to this theory, using concepts such as similarity and proximity it is possible to group information in a more useful and meaningful way[1].

4.4.3 Coding of information

Information in a display may be coded in many modes. The choice of a good coding simplifies the identification of information and reduces the time necessary to interpret it. This has a particular importance in the case of continuous feedback, as explained later in section 4.4.9. Information can be coded using text, shapes, colours, dimensions of objects, angles of lines, movement, etc. Grether & Baker (1972) and Smith & Mosier (1986) give indications on some of these modalities, as reported in Table 4.3.

[1] Gestalt theory was meant to have general applicability; its main tenets, however, were induced almost exclusively from observations on visual perception. Whatever their ultimate theoretical significance, these observations have been raised to the level of general principles. It is conventional to refer to them as Gestalt principles of perceptual organisation.

The overriding theme of the theory is that stimulation is perceived in organised orconfigurational terms (Gestalt in German means "configuration"). Patterns take precedence over elements and have properties that are not inherent in the elements themselves. One does not merely perceive dots; one perceives a dotted line. This notion is captured in a phrase often used to characterise Gestalt theory: "The whole is more than the sum of its parts."

Of the many principles of organisation that have been enunciated by various Gestalt theorists, the most general is referred to as Prägnanz. In effect, according to the principle ofPrägnanz, the particular perceptual configuration achieved, out of a myriad of potential configurations, will be as good as prevailing conditions permit. What constitutes a "good" configuration, or a poor one, is unfortunately not clearly specified, though several properties of good configurations can be listed, chief among them being simplicity, stability, regularity, symmetry, continuity, and unity. What happens when these properties of figures come into conflict is not specified, but should be possible to determine empirically.

The principle of closure often operates in the service ofPrägnanz; for example, a circular figure with small gaps in it will be seen as a complete or closed circle. Similarly, if a portion of the image of a figure falls on the blind spot of the retina, a complete figure often will still be perceived. Some distortions from good configurations may be so large as to preclude closure; in those cases, the figures may be a source of tension for the observer. (...) One Gestalt principle, that of common fate, depends on movement and is quite striking when observed. According to the principle of common fate, stimulus elements are likely to be perceived as a unit if they move together. An illustration of this principle is provided by a well-camouflaged object, such as a military vehicle; when stationary, the elements of the vehicle are integrated, through proximity, similarity, and so on, into patterns of background elements, and the object is difficult to detect. But it is easy to see it once it starts moving; with all of its elements moving in unison, the vehicle is readily perceived as a unitary figure, clearly segregated from its background. Movement is also at the heart of a set of observations of considerable significance in the historical development of Gestalt theory. These observations concern circumstances in which people perceive movement in the absence of actual physical motion of the stimulus. One familiar instance of this class of events is referred to as the phi phenomenon. In simplest form, the phi phenomenon can be demonstrated by successively turning two adjacent lights on and off. Given appropriate temporal and spatial relations between the two lights, an observer will perceive the first light as if it were moving from its location to that of the second light. The phi phenomenon is basic to the eye-catching displays used on theatre marquees and to cinematic and television presentations. The motion-picture screen, for example, presents a series of briefly flashed, still images; the movement people see is a creation of their own perceptual systems. (From Enc. Britannica)

Code	Max code steps	Recomm code steps	Evaluation	Comment
COLOUR				
Lights	10	3	Good	Location time short. Little space required. Good for qualitative coding. Larger alphabets obtainable by combining saturation and brightness with the colour code. Ambient illumination not critical factor
Surfaces	50	9	Good	Same as above, except ambient illumination must be controlled. Has broad application
SHAPES				
Numerals and letters	unlimited		Good	Especially for identification. Uses little space
Geometric	15	5	Fair	Particularly useful for symbolic representations. Some shapes may be more difficult to disseminate
Pictorial	30	10	Good	Allows direct association. Requires high display resolution
MAGNITUDE				
Area	6	3	Fair	Requires large space. Easy to find information
Length	6	3	Fair	Requires large space. Easy to find information
Luminance	4	2	Poor	May interfere with other signals. Ambient
Stereo-depth	4	2	Poor	Limited population of users
Angle of inclination	24	12	Good	Particularly good for quantitative evaluations, but applications are limited to round instruments, such as dials and clocks
Flash rate	5	2	Fair	Difficult to differentiate between flash rates. Good for attracting attention

Table 4.3. Coding modes

It should be clear that these guidelines have to be adapted to any specific application, since it is not possible to formulate general rules. Analytical methods such as task analysis, link analysis and simulation can help in the evaluation of the most suitable coding type in each case (Foley, 1982). The project can then be tested through experiments, evaluating the performance of the user, in same case evaluating his / her work load. Studying eye movement is also a well known technique (Kelley, 1968). In the specific case of representation of numerical data, Table 4.4 gives examples of techniques traditionally used, together with notes about their use.

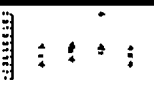
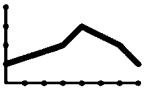
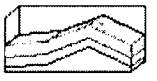
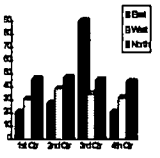
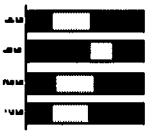
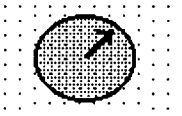
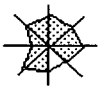
Graphical Technique	Example	Usage Notes
Scatterplots		Shows the correlation between 2 variables, or the distribution of points in a 2-dimensional space
Line Graphs or Curves		Shows how 2 continuous variables are related to each other. This technique is often used for displaying the changes of a variable over time. In this case time is generally shown on the horizontal axes. A third variable can be added by using different colours for each line.
Area, Band, Strata, or Surface Charts		Used to show how different line graphs or curves represent the parts of a whole. Bands of different colours represent the contribution for each category. It is generally preferred to draw the variables that change less on the bottom, in order to avoid a too irregular display, more difficult to read.
Bar Graphs, Column Charts, or Histograms		Used to show the value of a single variable over multiple separate entities, or to show a variable sampled at regular time intervals.
Stacked or segmented Bars or Columns		Similar to area, band, strata or surface charts, used when a number of bars represent the portions of a whole
Pie Charts		Show the relative distribution of data that form a whole (for example percentages). Often bar or column charts allow a more accurate interpretation of data.
Simulated Meters		Used to show one value of a continuous variable. In case more than one simulated meters have to be compared, it is often more effective the use of other techniques, such as column charts or line graphs.
Star, Circular, or Pattern Charts		Used to show the values of continuous variables on entities that are related. Values are drawn on spokes starting from the centre. With an appropriate scaling of values, normal values build a regular polygon. For this reason this technique allows a easy detection of values that are different from the normal. However this technique is not suitable for accurate comparison of data.

Table 4-4. Popular graphical techniques for representing numerical data.

4.4.4 Images

It is not clear whether images with full colours and full details are always desirable (B.Christie, 1991). Full colour pictures may be more attractive and may give more "feeling" or atmosphere, but it has been proven that information given through schematic drawings, such as cartoons, in some cases are remembered for a longer time. Champness & Ikhlef (1982) showed to two groups of subjects the same picture representing a house, and generated with two different graphic systems. The first image was of photographic type, with colours, and full of details. The second one was an image giving only schematic contours, obtained from a processing of the first image with a graphic package (Alpha geometric). The subjects were asked to remember how many windows the house had. People who saw the schematic image remembered with more precision. It appears that in certain cases a full detailed colour image may inhibit the comprehension of information, and that schematic drawings may be superior in showing structural details. This experiment confirms the conclusions by Millis (1982) on comprehension of cartoons, caricatures and pictorial metaphors. This may be explained in terms of *schema theory*, which suggests that schematic drawings and cartoons are closer to the way the brain codifies objects in *canonical form* (Hochberg, 1972), a type of simplified concept.

4.4.5 Animation

The use of animation has been until now limited to specific areas such as entertainment (films, video games etc.), education, training and presentations. Animation until recently a an expensive method to show information, and was therefore used only in the fields in which the same sequence is shown many times. Mills (1982) suggests that animated sequences may be well suited to the field of problem-solving, where interactive sequences can show the consequences of choices made by the user. Mills also distinguishes between *motion* and *sequence*, and affirms that a series of still images may give the perception of a sequence, and movement is not always necessary. Animated sequences are particularly useful to show cause and effect, especially if it is possible to jump to the desired sequence without the limitations of the previous techniques (linear video). While the capability of animated images is superior in showing the consequences of a choice, a static image sometimes may be more efficient for giving information which requires a certain time to be processed by the brain (for example when evaluating distances on a map). Animation is therefore a way to represent information that may have great potential, if used together with other techniques.

4.4.6 Mixed presentation forms

In mixed forms of presentation of information the problem from the human factor point of view is the integration of various forms in a way which is effective and easy to understand way, i.e. finding a consistent scheme. Which is the best way to show a combination of graphics, images, text and

animation? It looks like even obvious situations have to be studied with care. To cite a curious example, Ellis & Miller (1981) found that in the field of advertisement, right-handed people prefer pictures on the left of the text, while left-handed people do not express particular preferences.

4.4.7 Multimedia

The term *multimedia*, or more explicitly *interactive multimedia* defines any computer delivery system which allows users to access, control and combine different types of media, such as text, sound, images, video, computer graphics and animation. The most common applications include training programs, video games, electronic encyclopaedias, and tourist guides. "It moves the role of the user from observer to participant" (Encyclopaedia Britannica). The development of multimedia was also determined by the trend of the computer industry towards new fields, since the market for traditional types of computers (whose use is mostly calculation and data processing) was saturated. However, the diffusion of the CD-ROM, where the high volume of data required by multimedia applications is conveniently stored, made this technique enter gradually the standard features of graphic user interfaces, and it is now considered as a normal component of a computer system.

4.4.8 Virtual Reality

Virtual reality (VR) is the term given to the provision of artificial computer environments that a user can view or inhabit in some interactive manner. VR uses three-dimensional techniques of modelling and simulation to generate sensorial feedback as a result of movements and actions from the user. Feedback is generally visual and acoustic, but also tactile. In the case of *full immersion*, VR uses devices such as goggles, headsets, gloves and body suits, but the VR techniques are also useful to generate interactive pictures in real-time to be shown on traditional screens, in this case the use of pre-stored information is not possible. It is therefore an extension of the animation techniques, intended as a schematisation and simplification of reality (see previous section on animation).

4.4.9 Human information processing

In the next sections some aspects of human information processing are briefly given, since they are of practical use in the design of visual displays. Studies in experimental psychology on *integral and separable dimensions* help the choice of combination of graphic modalities for multidimensional and mono dimensional displays. A knowledge of some aspects of the early vision processes help in the choice of the metaphor to use in displays. Finally, knowing the reaction time of the human cognitive processes allows the calculation of the maximum delay the visual feedback is allowed to exhibit, in case the display is used to give a continuous feedback on user actions.

Integral and separable dimensions

Some visual dimensions are perceived as a whole (for example, hue and saturation of an object), while other are easily separable (for example, colour and size) (Garner, 1970; Kemler Nelson, 1993). There are differences in perception between children and adults (Smith, 1984). At first children see objects holistically, unable to pay selective attention to some characteristics, and ignore others. Gradually, some perceptual dimensions, such as colour and size, became separable, and can be used for categorising objects. However other dimensions, such as hue and saturation, remain integral (although this may be caused by difficulties in separation, rather than by an absolute inseparability) (Harnad, 1987). The issue is relevant when multidimensional displays have to be shown (for example for giving information at the same time about volume and pitch of a voice production). In such a case, the two dimensions chosen require to be separable (for example colour and location). Integral dimensions can be used to give some sort of redundant information in mono-dimensional display, to reinforce the information shown.

Early vision

It is useful to mention some aspect of the complex mechanism of visual perception, in which features automatically extracted from a scene are assembled into objects (Treisman, 1986). An hypothetical model of early stages in visual perception considers a pre-filter which codifies some simple and useful properties from a scene in a number of "feature maps" such as colour, orientation, size and stereo distance. This stage locates the objects in a spatial map. An "attention spotlight" mechanism selects and integrates the features in a certain location (see Figure 4.2).

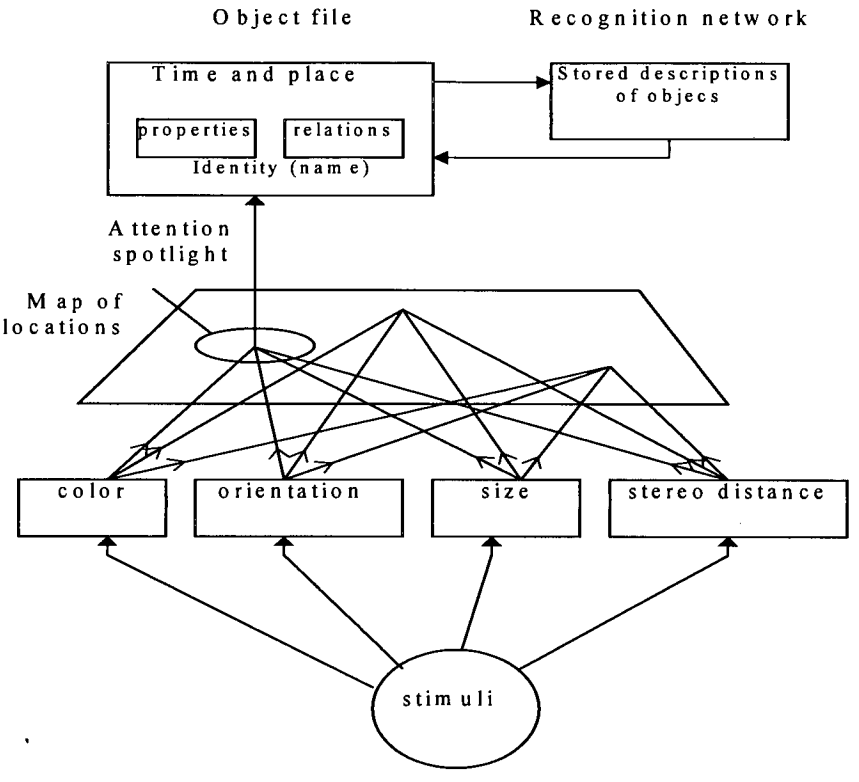


Figure 4.2 Hypothetical model of the early stages in visual perception

This information, in the subsequent stages, is compared with descriptions stored in a recognition network. From these studies it results that a "knowledge of the world" speeds up perception and make it less subject to errors, a consideration that has to be taken into account when choosing the method to display information.

Colours

Some consideration can be added to the guidelines about colours seen in the previous sections. Studies about colour perception (for example Brou at al., 1986) show that the same colour is perceived differently depending on the colours surrounding it. For example, a small grey area on a blue background is perceived as "warmer" (more reddish) than the same grey area on a red background. This suggests that displays conveying information by a change in colour should keep the surrounding areas with a constant colour, to avoid incorrect evaluations.

Feedback-control mechanism of human behaviour

The visual interface in some applications gives a continuous feedback of the users actions. An arcade game display, for example, shows the effect of the user's actions, and the user corrects the actions depending on the information on the video-game display. The system consisting of the human and the arcade game work in a close-loop mode. It is important to know human response times in order to define the maximum response time of the arcade game to make the closed-loop work effectively. If the response time of the arcade game is too long, the whole system becomes unstable.

Defining the maximum allowable delay of the arcade game needs information about the way that humans process information. A simplification of the *human processor* by Card *et. al.* (1983) gives enough information to calculate the human response time. In this view, the human contains three processors, a Perceptual processor, a Cognitive processor and a Motor processor, all operating in "pipelined parallel" (see Figure 4.3). A person can thus read a word while saying the previously read one. In the figure the model is shown in the process of receiving a visual stimuli and reacting to that via a motor action (in this case, pressing a button). The typical processing times are also shown.

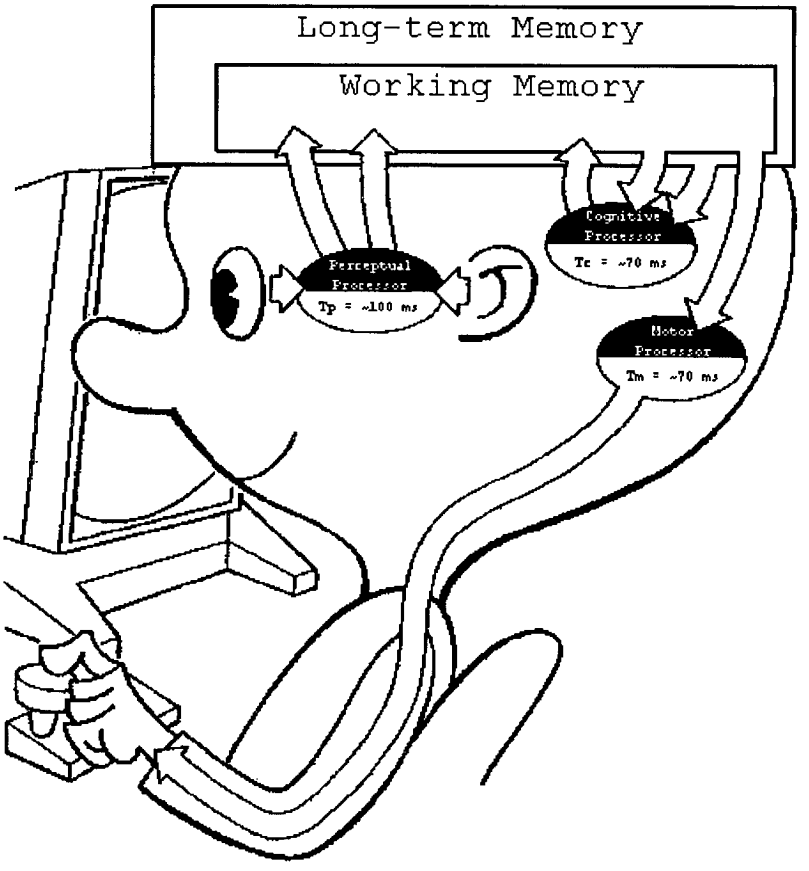


Figure 4.3 The Model Human Processor (Card, 1984)

The figure shows two memories, the *working memory,* that works as a cache memory, where recently experienced and currently active information are accessible; and the *long-term memory*, in which the users hold their general store of knowledge. Current opinion in psychology does not consider these two memories as a separate set of storage cells; instead the long-term memory is thought as a graph of semantically linked nodes, and the working memory is a small subset of these nodes, which are active at a given moment.

The total time deduced from the model is:

Perceive stimulus	Tp =	100 [50 ~ 200] ms.
Decide to respond	Tc =	70 [25 ~ 170] ms.
Respond	Tm =	70 [30 ~ 100] ms.
	Total	240 [105 ~ 470] ms.

The human processor can thus perceive, decide and respond to an event in approx. 240 ms. It has to be added that the perceptual processor treats stimuli that happen in a time less than its processing time as a single, averaged stimuli. What is the maximum time the external device (the arcade game) can take to show the result of an action performed by the user in order not to make the human processor take a wrong decision? It depends on the type and speed of feedback and the action required, with the worst case in the range 105-470 ms. as shown above. In most cases, when the quickest correction of an action is not required, techniques normally used in servo-control systems can be applied to avoid over-reactions from the user. An appropriate low-pass filter on the feedback response time is generally an adequate solution.

4.5 A comparison between visual interface guidelines and visual feedback for the hearing impaired

The previous section has presented the requirements that a visual display has to satisfy in order to be clear to understand and easy to control. Coding method, amount of information to display, placement of information, response time, all these factors affect the effectiveness of a visual display.

Knowledge of these topics provides an essential background information for building new visual feedback for hearing-impaired speech rehabilitation. But it would be interesting to investigate if these guidelines were followed in the design of the existing visual feedback.

Table 4.1 in Section 4.2 attributed some problems to poor human factors and system design. The first of these problems was "lack of motivation". Therapists explained that their clients would like much more variety in their visual feedback, with the use of more "visual", "multimedia" and "arcade-game" type of displays. The introduction to visual interface in Section 4.4 showed a wide range of modes and techniques, such as virtual reality, animation, multimedia, that don't seem used enough (or at all) in visual feedback for speech rehabilitation.

Another problem was "difficult navigation" in the system. This may depend on a poor or inconsistent placement of controls such as menus and button, or overcrowded displays where the necessary information is difficult to locate.

About "multidimensional displays", without entering the discussion if they are appropriate or not[1], some inconsistency was noticed in the literature, as discussed in Section 4.2. To make things more complicated, the debate on separable and integral displays carried out by experimental psychologists is still open, exhibiting contrasting findings (compare for example Melara et al., 1990, with Kemler Nelson, 1993).

To summarise, some improvement in usability of the system can be achieved by following guidelines on visual display design, and possibly by selecting more carefully the visual dimension for multidimensional displays. But it looks like the most promising issue for improving the effectiveness of current visual feedback is experimenting with the vast range of coding methods and graphic techniques, also adding some of them which have never been used before. The systems today are similar in the choice of the coding methods and graphic techniques being used. Intensity is generally represented by an object changing its size (for example a balloon that becomes bigger with the increase of the speech intensity), pitch is represented by an object that moves up and down, vowel

[1] Regarding this topic see for example Powel & Maassen (1987), and Watanabe, Ueda & Shigenaga (1985).

quality is represented by an object that moves in a bi-dimensional space. Apart from studies recently started on visual perception in deaf people (Gallaudet Institute, 1994) of which results are not yet available, it appears that until now graphic displays were not based on research explicitly focused on deaf people. Most displays were realised using "common sense" principles, and the effectiveness has then been evaluated (sometimes in a quite extensive way, as in the case of the VisiPitch and the SpeechViewer). It has still to be proven that the feedback modalities now in use are really the most effective.

4.6 Needs for experiments for finding intuitive feedback

Two important issues are open:

- Are the visual feedback approaches currently in use the most effective?
- How should new graphic techniques, such as multimedia and virtual reality be used to give a better visual feedback?

It was decided to run a series of experiments in order to evaluate the intuitiveness and enjoyability of a range of coding methods and graphic modalities, both those used traditionally, and new ones never tried before in the field of deaf voice rehabilitation. These experiments carry-out the third step of the "new approach" explained in Section 4.3. The experiments do not try to give a final answer to the open questions stated above, but to define some point that will be useful for building an initial system, which will be in turn tested for effectiveness with the help of a number of patients and therapists. The design and implementation of the experiments is carried-out in Chapter 5.

4.7 Conclusions

Hearing-impaired speakers and speech therapists often encounter problems in the use of speech rehabilitation systems. An analysis of the problems that therapists and users encounter during speech rehabilitation sessions shows that there is room for great improvement. It appears from the studies carried out until now that visual feedback used in the majority of speech rehabilitation systems for the hearing impaired were "imposed" on the users, who had to learn how to use them. This was determined by the small number of theoretical or empirical studies capable of providing information on the relation between hearing and vision. A new approach in designing visual feedback for the rehabilitation of the hearing-impaired was then proposed. In this approach, an intuitive link between speech features and visual dimensions is investigated by means of experiments. These experiments are carried-out in the next Chapter.

CHAPTER 5

An Investigation of Visual Feedback Responses for Hearing-Impaired Speakers

5.1 Introduction

As discussed in the previous Chapters, visual feedback for speech rehabilitation needs improvement for the following reasons: 1) therapists have some complaints about present systems because they may give negative reinforcement, cause frustration, and may be inaccurate and difficult to understand; 2) current visual feedback approaches are based on practical approaches rather than on extensive research. The issue of how to best represent a particular speech feature is still open. The first experiment described in this chapter approaches the second of the points mentioned above, and describes a novel experimental method where, instead of proposing different visual representations for the various speech features (such as loudness, pitch, vowel quality etc.) and assessing which work best, the various visual stimuli are shown to the subject without specifying the associated speech feature. In this way an intuitive connection between visual stimuli and speech features (the sound produced by the subject) can be characterised. The goal of the experiment is to identify the best associations for visual stimuli and speech features within the group of hearing-impaired subjects. The visual stimuli for each association is then used in a visual feedback scheme for that speech feature. A comparison with responses of normal hearing subjects is also attempted.

A second experiment compares the subject's preferences on simulated 3D graphics realised with virtual reality techniques and shown on a conventional computer screen, with a "full immersion" version of the same graphics, displayed using a 3D headset.

A third experiment uses *multimedia* techniques to assess the subject's acceptability of and motivation towards this methodology.

5.2 Experiment 1

5.2.1 Methods and Procedures

The goal of this experiment is to study which features of the voice are intuitively linked to different graphic modalities.

5.2.1.1 Selection of stimuli modalities

The selection of visual stimuli using different graphic modalities followed these rules:

1. Build dynamically changing stimuli, to simulate a display giving real-time feedback.
2. Consider as a starting point the wide range of coding methods and graphical techniques to display information suggested by visual display guidelines.
3. Discard those not suitable for displaying speech features.
4. Include those already used in therapy in order to assess their association with their relative speech features.
5. Keep stimuli simple in order not to add extraneous elements which are difficult to evaluate.

Coding methods and graphical techniques for displaying numerical data were presented in Section 4.4.3. Here they are listed again:

Coding methods

- Magnitude:
  - ◊ Area
  - ◊ Length
  - ◊ Luminance
  - ◊ Stereo-depth
  - ◊ Angle of inclination
  - ◊ Flash Rate
- Shape
- Colour

Graphical techniques

- Scatterplots
- Line graphs or curves
- Area, band, strata ...
- Bar graphs, column ...
- Pie charts
- Pattern charts

Other more complex graphical techniques, such as *virtual reality* and *video*, are addressed in Experiments 2 and 3.

The following section reviews the coding methods and graphical modalities listed above, in order to decide which are suitable for including in the visual stimuli for the experiment. The decision considers previously published uses in deaf speech therapy (see Section 3.3), and recommendations from visual interface guidelines (Grether & Baker, 1972; Tullis, 1988; Smith & Mosier, 1986).

Area

Visual interface guidelines consider *area* (as a method for representing mono-dimensional data in the form of screen objects varying in size) only for applications where the displays are not crowded. When using area as a quantitative measure, no more than 6 code steps are recommended. For size coding, a larger symbol should be at least 1.5 times the height and width of the next smaller symbol.

In many voice rehabilitation systems area is used as a feedback for speech loudness. An object generally with rounded shape, such as a balloon or a whale, becomes bigger as loudness of the voice increases, often with continuously variable size (infinite or high number of code steps). In such an application the recommended maximum 6 code step is not a limiting factor (compare with a six bar VU-meter on a tape recorder, giving an adequate loudness range). However, if used for relative measurement, as is generally the case in therapy, a continuously variable area gives, in fact, a more adequate feedback. The fact that the display may use a large area of the screen is not a limiting factor either, since in this particular application the main area of the screen should be reserved for the feedback itself in order to focus the user's attention on the feature of interest. Stimuli using this technique will therefore be included in the experiments, in order to assess the association between size and loudness already used in speech therapy, and also to investigate other possible links with other speech features.

Length

Visual interface guidelines suggest that long lines will add clutter to a display, but may be useful for special applications. No more than 6 code steps are recommended. In deaf speech therapy the length of lines or other thin objects is used inconsistently. In some systems the length of a vertical line varies as a display of pitch, in other systems the length of a vertical or horizontal line varies as a display of voice intensity. Length on the horizontal axis is also used to represent the duration of an event. The concept of "height" of a sound as a way to represent its fundamental frequency is familiar to most normal-hearing people (for example, tone controls on audio systems are often called "low" and "high" frequency controls). Because of that the common association of pitch with length of a vertical line (or vertical position of an object) is widely used. However it may be different for hearing-impaired people. The length of an horizontal line for representing voice level is a convenient display option for calibration purposes, however it is not widely used as a method for feedback, being generally replaced by other types of feedback, such as area. Since the variation of length is used in speech rehabilitation systems as feedback for different speech features, stimuli featuring change of length will be included in the experiment to characterise intuitive connections with speech features.

Luminance

Visual interface guidelines advise against the use of luminance as a coding method because of its dependence on ambient illumination. In any case, no more than 4 code steps are recommended. In deaf speech therapy, the only reported cases of use of luminance in speech rehabilitation systems are as a feedback for voice level. This is unusual since generally voice level is represented by the size of an object. It is understandable how the ambient illumination can influence the correct interpretation of

information coded using luminance. However, one paper (Povell & Maassen, 1987) suggests that the poor visibility of an object which is too dark, and the excessive brightness of an object which is too bright, may be advantageously used to push the user to "stay in the middle". Therefore, in cases where the importance of the information is relative rather than absolute, this modality may be interesting. Stimuli using change of luminance will be included in the experiment to assess the association with loudness of this modality which has been used very rarely in speech rehabilitation systems, and also to investigate if there are links with other speech features.

Stereo-depth

Visual interface guidelines advice against this modality because of the limited population of possible users. There are no reported cases of use in deaf speech therapy. Stereo-depth, giving the perception of distance of an object, is one of the visual features that is pre-processed by early stages of vision (Treisman, 1986), together with colour, inclination angle, dimensions, and some features of shape. It not clear how this may influence the efficacy of visual feedback. Use of a 3D headset may be expensive and invasive, but may be interesting to evaluate the effect of a simulated stereo-depth on a normal 2D screen. Therefore stimuli giving the impression of an object coming closer to the subject and going far away will be included in the experiment.

Angle of inclination

The human vision apparatus is particularly sensitive to inclination angle (Treisman, 1986). More than 20 code steps are normally achieved. When used in simulated meters, an indicator changing its angle of inclination is suitable for displaying the value of a single continuous variable. In case of multiple variables, visual interface guidelines recommend consideration of other techniques (bar, columns charts or line graphs) rather than a series of simulated meters. In deaf speech therapy, apart from "S" indicators (see Section 3.3.5) there are no reported cases of visual feedback using this modality. Arrows with different orientation are sometimes used to indicate changes in pitch, or stress (however not in real-time feedback). It is difficult to define an intuitive link between a speech feature and this modality. Nevertheless, since the human vision system is sensitive to angle of inclination, it is worth investigating its use as a direct feedback method. Therefore stimuli using changes in angle of inclination will be included in the experiment.

Flash Rate

Visual interface guidelines recommend the use of flash rate (or blink code) when urgent attention by the user is required. Suggested rates are between 2 and 5 Hz, and although it may be possible to use a maximum of 4 code steps, not more than 2 are advised (blinking versus non-blinking). In deaf speech therapy, blinking lights are sometimes used for signalling excessive voice level, or as a reward in voice controlled video games. There are no reported cases in which this modality is used as a direct visual feedback method. In some new vibro-tactile feedback techniques, a device stimulates the skin at a sub-multiple of the voice fundamental frequency, in order to convey pitch to the hearing-impaired user (Bernstein, 1995). It is worth investigating whether the same technique works with flashing lights[1], therefore stimuli using flash rate will be included in the experiment.

Shape

Visual interface guidelines suggest that shapes are useful for their capability to allow direct associations, and 30 code steps are achievable. Shape codes can be mnemonic in form, and in any case their interpretation will generally rely on learned association as well as immediate perception. Furthermore, possible existing standards need to be taken into account. In deaf speech therapy, abstract types of shapes were used many years ago in multidimensional displays to give a complete feedback of speech sounds (the "Spectral Movie", Nickerson & Stevens, 1973) and Lissajous[2] figures were also used to display spectral characteristics of speech. Recently only a few applications using shape as a feedback mechanism have been reported (such as the "Visual Hear", Reynolds, 1993). A different case, featured by some new systems, is the use of vocal tract shapes to display articulators not visible externally. There is not enough information to evaluate if the use of abstract shapes can be effective in visual feedback. Therefore, it is worth investigating this coding method by including in the experiment some stimuli using simple changes of shape.

Colour

As stated by visual interface guidelines and studies on the vision mechanism (Treisman, 1986; Brou et al., 1986), colours are a powerful method for coding information. They allow fast location of information, and they can be used for grouping and enhancing information. With suitable

[1] It was decided to include this modality in the experiments after an exchange of ideas between Lynne Bernstein (Gallaudet University for the Deaf, Washington DC) and the author.

[2] Jules Antoine Lissajous (1822-1880) was interested in waves and developed an optical method for studying vibrations. At first he studied waves produced by a tuning fork in contact with water. In 1855 he described a way of studying acoustic vibrations by reflecting a light beam from a mirror attached to a vibrating object onto a screen. He obtained Lissajous figures by successively reflecting

combinations of hue and luminance, 50 code steps are achievable. If a relative value of a variable has to be displayed, it is preferable to use a gradual change of saturation instead of hue. Relative comparisons of colours are difficult, and they are possible only when they are shown at the same time, and, in general, absolute judgement is difficult. Factors such as ambient illumination and display background can influence the colour perception. When choosing colours to code information it should be considered that they are often associated with conventions: for example red is often associated with danger, yellow with caution, green with normal status. White is neutral. In deaf speech therapy colours are used in many, often inconsistent, ways. In case of feedback for voice level, some systems use black, some others grey to represent silence, green or grey for low level, green, blue or yellow for normal level, red or white for excessive level. In all of these cases colours are used categorically, and chosen in an arbitrary way. There are no reported cases of feedback in which a particular colour was chosen because it makes sense for a particular feature of speech[1]. Furthermore, there are numerous studies on *synesthetic*[2] links between visual dimensions and auditory dimensions (for example Melara,

---

light from mirrors on two tuning forks vibrating at right angles. The curves are only seen because of persistence of vision in the human eye.

[1] Watanabe (1985, 1995) uses colours to visualise speech, as discussed already in Section 3.3.5, but the choice of colours is based on a principle of "uniqueness" rather than "natural similarity" with speech features. In case of vowels, the Red, Green and Blue (RGB) components of colours are calculated extracting the first three formants (F_1 .. F_3) and applying the following equations:

$$\text{Red} = k\,(5\,F_1 / F_3)$$
$$\text{Green} = k\,(3\,F_3 / 5\,F_2)$$
$$\text{Blue} = k\,(F_2 / 3\,F_1)$$

where k is a scale factor.

This results in the following representative colours for the five Japanese vowels /a, e, i, o, u/:

/a/: orange-yellow-greenish yellow with high saturation
/e/: magenta with low saturation
/i/: blue with very high saturation
/o/: green-cyanic green with high saturation
/u/: cyan with middle saturation.

[2]Synesthesia (Greek, syn = together + aisthesis = perception) is the involuntary physical experience of a cross-modal association. That is, the stimulation of one sensory modality reliably causes a perception in one or more different senses. Perhaps the most famous family case is that of the Russian novelist Valdimir Nabokov. When, as a toddler, he complained to his mother than the letter colours on his wooden alphabet blocks were "all wrong," she understood the conflict he experienced between the colour of the painted letters and his lexically-induced synesthetic colours. [...] As early as 1704, Sir Isaac Newton struggled to devise mathematical formulae to equate the vibration of sound waves to a corresponding wavelength of light. [...] Goethe noted colour correspondences in his 1810 work, Zur Farbenlehre. [...] The nineteenth century saw an alchemic zeal in the search for universal correspondences and a presumed algorithm for translating one sense into another. [...] Synesthesia attracted serious attention in art, music, literature, linguistics, natural philosophy, and theosophy. Two books were published: 'L'Audition Colore' by Suarez de Mendoza in 1890, and 'Das Farbenhren und der synsthetische Faktor der Wahrnehmung' by Argelander in 1927. Most accounts emphasised coloured hearing, the most common form of synesthesia. [...] Vasilly Kandinsky (1866-1944) had

1989), but these have not resulted in applications for hearing-impaired speech. Therefore simple stimuli using an area changing colour will be included in the experiment.

Scatterplots

Scatterplots, where data are plotted as points in a two-dimensional graph, are useful for showing distribution of points in space, or how two variables are correlated. Visual interface guidelines suggest that significant data points may be highlighted by colours, blinking, shape coding or other means. Consistency of axes should be respected if more than one plot is displayed. In deaf speech therapy, scatterplots (fixed or dynamic), are often used to represent vowels on a two-dimensional plane, plotting F_1 vs. F_2 or variations of this (see for example Arends, 1993). Reference vowels are "clouds" of points on which attempts by the talker to produce the desired vowel are superimposed as distinctive dots. Such a display however lacks intuitiveness, and the user has to learn how to interpret it. Stimuli using this modality will be included in the experiment, in order to assess whether they can suggest some association with speech features that can be described with two dimensions, or with one dimension plus a time axis. However, to avoid showing a complex graphic that may bias the experiment, the stimuli will be simplified, showing a simple object changing position on a plane.

Line graphs or curves

Line graphs and curves are suitable for representing relations between two continuous variables, particularly to display data changes over time. Line graphs are a special form of curve, so visual interface guidelines recommendations for both modalities are the same. Computer generated curves can be used to show dynamic data change. Several curves can be compared with others, but it is advisable not to show more than four curves at a time. In deaf speech therapy, curves are often used to display pitch or loudness change (vertical axis) over time (horizontal axis). Curves offer a clear method for displaying pitch and loudness over time. However, since the same technique is used for different purposes, this may be confusing for the user (see for example Povel & Maassen, 1987 about the need for uniqueness of appearance). Furthermore, there is a need for some more intuitive method to display features such as pitch and loudness, especially in the early phases of speech rehabilitation (awareness modules). For loudness, an alternative often used is the change in dimensions of an object, while for pitch, there are generally no alternatives used. The consideration made above for the

---

perhaps the deepest sympathy for sensory fusion, both synesthetic and as an artistic idea. He explored harmonious relationships between sound and colour and used musical terms to describe his paintings, calling them "compositions" and "improvisations." His own 1912 opera, Der Gelbe Klang ("The Yellow Sound"), specified a compound mixture of colour, light, dance, and sound typical of the Gesamtkunstwerk. (from "Synesthesia: Phenomenology And Neuropsychology A Review of Current Knowledge" by Richard E. Cytowic, 1995).

association between pitch and vertical position of an object also applies here. Stimuli using dynamic line graphs and curves will therefore be included in the experiment, to assess the speech features are intuitively associated with them.

Motion, speed

Visual interface guidelines recommend that between 2 and 10 degrees of motion might be distinguished, in applications where this method is an appropriate means of display coding. In deaf speech therapy, motion is used in drills for voice onset time, where an harsh attack causes a quick jump of a pointer, while a softer attack causes a slower and shorter movement. Psychology literature suggests other uses of speed. Relaxation is associated with calm, slow movements. Speed is associated with tension. Tension tends to increase a speaker's pitch level (Laver, 1980). It is therefore worth investigating if speed can indirectly influence a speaker's pitch. Therefore, stimuli using motion and speed will be included in the experiment.

Area, band, strata or surface charts

Visual interface guidelines recommend the use of area, band, strata or surface charts when several line graphs or curves represent all the portions of a whole. There are no reported cases of visual feedback using this modality in deaf speech therapy. Indeed this modality does not seem appropriate for speech feedback, since it represents several variables with the same method, while the goal for visual feedback is to try to characterise different speech features in distinct ways. Therefore stimuli using these modalities will not be included in the experiment.

Bar graphs, columns

Visual interface guidelines recommend the use of bar graphs and columns to compare a single measure across a set of several entities, or for a variable sampled at discrete intervals. For expert users the overall pattern of a bar graph may serve as a diagnostic function beyond the comparison of individual bars. For example, if multiple bars are used to show different components of a system, users may learn characteristic "profiles" of the bars which indicate system status. In deaf speech therapy, bar graphs representing the energy spectrum are often used for feedback of vowels and consonants. However, reading a spectral display is not an easy task (Cole & Zue, 1979) and it is very difficult to try to correct a vowel relying only on feedback from a spectrogram. On the other hand, the same display can be useful as an "awareness" module in the early stages, just because it moves in various ways depending on loudness and spectral contents at the same time. At such a stage it is not important to be able to detect "what causes what", and the most important thing is to provide some

enjoyable tool to motivate users to experiment with their own voice. However, since this modality cannot be used as a feedback for a single feature of speech, it will not be included in the experiment.

Stacked or segmented bar graphs or columns

According to the visual interface guidelines, for these techniques the same considerations made for bar graphs and columns are applicable. In deaf speech therapy there are no reported cases of visual feedback using this modality. As for bar graphs and columns, these modalities do not appear suitable as feedback because of the difficulty in correlating different speech dimensions with the variables controlling the graphics. Furthermore they cannot be used as a feedback for a single feature of speech. Therefore they will not be included in the experiment.

Pie charts

Visual interface guidelines recommend consideration of the use of pie charts only in special cases, to show the relative distribution of data of different categories. Generally a bar graph permits more accurate interpretation. In deaf speech therapy there are no reported cases of visual feedback using this modality. Since pie charts give another method of comparing variables of the same type, they are not suitable as a feedback for speech, where different speech features should be displayed with different visual dimensions. Therefore stimuli using this modality will not be included in the experiment.

Pattern charts

Visual interface guidelines suggest that pattern charts are suitable for detecting patterns, but not for detecting precise values or for accurate comparisons among values. In deaf speech therapy, some systems use pattern charts to show in a single display how a speaker's speech features are close to normal values. However this modality is not used as a direct feedback to the user, and it is intended to be read by the speech therapist. Since this modality is not suitable for displaying a single speech feature, stimuli using this modality will not be included in the experiment.

Realistic stimuli

A further group named *realistic stimuli* have been added to evaluate the response of the subjects to more complex (and possible more enjoyable) graphical techniques. The stimuli in this group exhibit a high degree of realism. These stimuli are an extension to the "speed" stimuli, with the difference that it is the viewer who is moving, instead of looking at objects moving.

The following table summarises the considerations and decisions made above.

Feature	Possible use	Published use	Included in experiment
Area	Yes	Yes, for loudness	Yes
Length	Yes	Yes, for different features	Yes
Luminance	Yes	Yes, for different features	Yes
Stereo-depth (distance)	Yes	No	Yes
Angle of inclination	Yes	Yes ("S" indicators)	Yes
Flash rate	Yes	No	Yes
Shape	Yes	Rarely	Yes
Colour	Yes	Yes, but categorically	Yes
Scatterplots (position)	Yes	In vowels	Yes
Line graphs or curves	Yes	Yes, for different features	Yes
Motion, speed	Yes	In some cases (voice onset)	Yes
Area, band, strata..	No: multiple variables	No	No
Bar graphs, column...	No: multiple variables	Rarely, for spectrum	No
Pie charts	No: multiple variables	No	No
Pattern charts	No: multiple variables	No	No
Realistic stimuli	Experimental	No	Yes

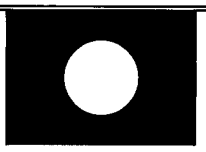
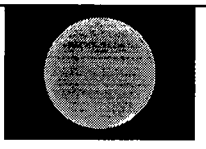
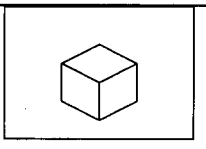
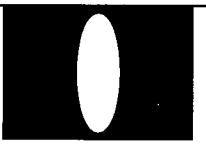
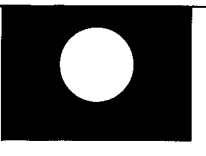
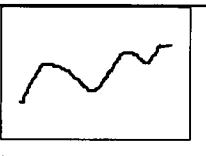
Table 5.1. Selection of stimuli to be included in the experiment.

5.2.1.2 Implementation of the stimuli

The visual stimuli have been generated with "3D Studio" software on a PC. This software allows design of 2D and 3D objects, and optionally animation. For animated objects, the process of "rendering" each frame may take a considerable amount of computation time. In this case all the frames are previously computed (the process may take several minutes or hours) and saved as separate images. The sequence of images is then played back with a separate piece of software, the *animation player.*

From the selection criteria described above, 12 groups of stimuli were defined, with one or more variants of the same type of stimuli in each group: Dimension (Area), Length, Luminance, Distance (Stereo-depth), Angle of inclination, Shapes, Colours, Flash rates, Line graphs and curves, Position and motion (scatterplots), Speed, "Realistic" stimuli.

The following Table 5.2 describes each stimulus.

#1: Area		A white circle in the middle of the screen, black background. The circle enlarges and shrinks at different speeds.
#2: Length		An horizontal line lengthens and shortens at different speeds. A vertical line lengthens and shortens at different speeds.
#3: Luminance		A white circle in the middle of the screen, black background. The circle becomes brighter and dimmer at different speeds.
#4: Distance (Stereo-depth)		A cube with a 3D appearance comes close to the viewer and goes father away, at different speeds.
#5: Angle of inclination		A line rotates 360 degrees clockwise and counter-clockwise around one of its ends, at different speeds.
#6: Shapes		An horizontal oval shape gradually changes into a vertical oval shape, and back, at different speeds. An horizontal toroidal shape gradually changes into a vertical toroidal shape, and back, at different speeds.
#7: Colours		A circle changes colour (hue) gradually, retaining the same saturation and brightness, at different speeds.
#8: Flash rates		A white circle on black background flashes at different rates.
#9: Line graphs, curves		A curve develops from left to right, following a few typical patterns for pitch contour or loudness feedback.

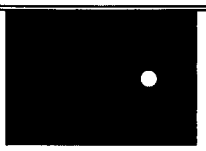
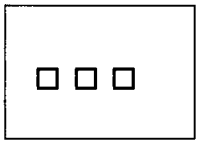
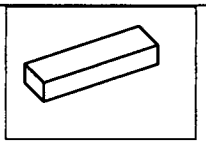
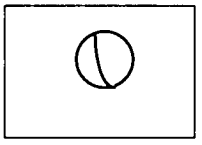
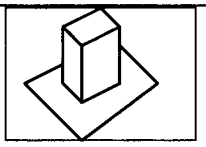
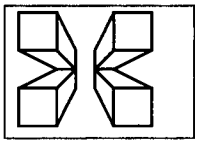
#10: Position, Motion	 	A small white circle on a black background moves periodically horizontally. As above but vertically. As above but diagonally. As above but moving following a sinusoidal function. A small cube appears, then another one on its side, then a third one
#11: Speed	 	A bar with a 3D appearance rotates around its centre at different speeds. A textured sphere with a 3D appearance rotates around its centre at different speeds. A hot air balloon flies between the clouds and the ground at different speeds.
#12: "Realistic" stimuli	 	The viewer has the impression of climbing and descending a parallelepiped, at different speeds. The viewer has the feeling of looking down from a helicopter, landing on a street surrounded by high buildings. The helicopter then takes off again. The scene is repeated at different speeds.

Table 5.2 The visual stimuli selected for Experiment 1

5.2.1.3 Subjects

The volunteer subjects for this research belong to two categories: hearing-impaired people and normal-hearing people. The hearing-impaired subjects were fifteen deaf people of both sexes: five children, age between eight and fourteen, and ten adults[1] aged over twenty, with deafness ranging between 50 (moderate-severe) and 90+ dB (profound) for the better ear. All subjects were undergoing speech therapy in schools for the deaf or speech rehabilitation laboratories.

The total number of fifteen hearing-impaired subjects is not very high for producing results with statistical significance. However it is very difficult in practice to recruit even this number of people. Some schools and teachers are very sceptical about the use of computers for speech rehabilitation, and some tend to teach sign language only to allow deaf people to properly communicate with each other, without any attempt to rehabilitate the speech and integrate them in normal-hearing society (Sacks

[1] During the experiments two out of the ten hearing-impaired adult subjects were not in the condition of producing any sound. The data from those sessions were removed from the results.

1989). The reasons for this policy are worthy of respect, but the result is that these schools declined to lend support for the research. The schools for the deaf that use computer assisted methods were very helpful, but nonetheless it took a long time for the therapists who assisted this research, to arrange for groups of at least two people at a time to attend the experiments. With children there was the issue of obtaining formal permission from the parents, and this resulted in only a small number of children (five) being available. On the other hand, there was no problem in finding normal-hearing subjects. In order to balance the experiment the same number of subjects was used: the number of normal-hearing subjects, of both sex, were ten adults aged over twenty, and five children, aged between four and nine, all with no knowledge of speech rehabilitation techniques.

5.2.1.4 Structure of individual trials

The experiments with hearing-impaired subjects took place in three different cities in the UK (Edinburgh, Glasgow and Southampton). The rooms chosen for the experiments were quiet and generally known by the subjects. One subject at the time was present. Apart from the subject and the researcher, the others present in the room were the teacher or speech therapist, and in a few cases a relative or friend of the subject. The experiments with normal-hearing people took place in a quiet room in the University of Edinburgh.

The subject sat in front of a 15" computer screen, at a distance of about 0.5 metres. For conditions of repeatability, the computer screen was calibrated in brightness, contrast and gamma correction with the "Kodak Precision Colour Management System" (Kodak, 1996) and checked against colour test cardboard references. Notice that this test takes into account ambient lighting levels, so that the image in the screen looks the same under different lighting conditions. Also it was checked that there were no light reflections on the screen. A S-VHS videotape recorder was used to record the subject's speech together with the visual stimuli shown on the computer screen. The subject's face was recorded on video as well, in a small area on the side of the visual stimuli, using a mirror (see Figure 5.1) This was used to check if the subjects were actually looking at the screen, if they were distracted or bored, and so on.

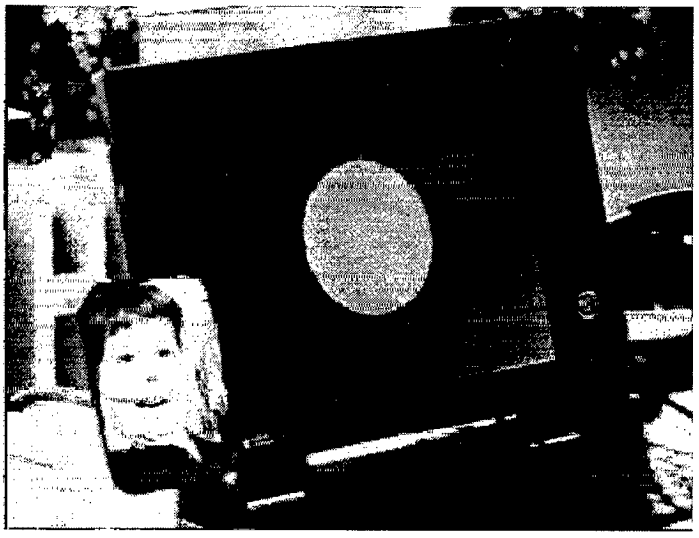


Figure 5.1 The mirror on the side of the computer screen, with the reflection of the child attending the experiment. This picture was extracted from one of the experiment videotapes.

The voice of each subject was recorded using a condenser super-cardioid flat microphone (Crown model PCC160) placed on the table at the side of the computer screen. This microphone, excellent in linearity and off-axis noise rejection, was selected in preference to head-worn microphones since it is non-invasive and designed for discretion. This helped to avoid subjects feeling uncomfortable when speaking and producing unusual noises.

The stimuli were presented in a random sequence, using the following method: The stimuli group was chosen randomly. If the chosen group consisted of more than one stimulus, they were shown together, in a random order. The next group was then randomly chosen, excluding from the choice the group already shown, and so on. Between one stimulus and the next, there was a pause of about five seconds. At the end of all the stimuli in all groups, the same sequence of groups and stimuli was shown again a second time, to check for stability of choice.

5.2.1.5 Priming of the subjects

Subjects were given the following instructions: "On the screen in front of you will see different moving images. While watching the images, try to 'comment' on them with your voice in the way that you prefer, that you feel most appropriate to 'match' the images. Use your voice as you like. I'm not trying to see if you are doing right or wrong, I'm just interested in listening to your voice and seeing what you think is the sound of the images that are on the screen".

Care was taken not to suggest examples of audible noises which may bias the subject in favour of any kind of speech production. This "complete freedom", was sometimes cause of undesirable effects, such as attempts by subjects to describe the stimuli verbally.

The stimuli were then shown for the first time. At the end of the first showing, the stimuli were presented again in the same sequence as before. Only the data recorded in the second show were taken into account in the experiment. It was immediately obvious that subjects needed to "break the ice" with the experiment (task habituation), with the new stimuli and with the understandable shyness due to the presence of an extraneous person and the uncomfortable knowledge that they were taking part in an experiment. Most of them felt more relaxed after the first showing. During the first showing, if the subjects were not producing any sound with their voice, they were encouraged with gestures to do it. Children behaved differently. They generally were immediately confident and curious, while during the second pass they tended to be more bored and uninterested. In some few cases, children did not complete the second showing because they remembered they had "seen it already". For these reasons, it was decided to take into account only the data recorded in the first showing, with these children.

5.2.1.6 Interviews

At the end of the session subjects were asked to rate on a form how much they liked the different stimuli. The form had five scores for each stimuli, "strongly disliked", "disliked", "neutral", "liked", "strongly liked", with accompanying pictures showing "smiling faces" or "unhappy faces" to make things clearer and more attractive for children. There was also a space for comments.

5.2.1.7 Evaluation methodology

The results were evaluated with the help of three independent professional assessors who were expert phoneticians. They watched the videotapes in groups of two (one of the experts and the author), and produced three sets of evaluation data. In case the evaluations mismatched in some parts, the relevant part of the video tapes was evaluated again together with the experts who disagreed, in order to have a discussion and a final agreement. All conflicting cases were solved in this way. The videotapes were not evaluated with all the experts together at the same time because of the difficulty in arranging their availability for several hours on the same day.

In the evaluation of the results the following issues were estimated and scored with a value from 0 to 10:

- Interest / enjoyment
- Motivation to speak
- Significant changes in loudness
- Significant changes in pitch
- Significant changes in vowel quality

Details of these issues are discussed below:

Interest / enjoyment

Interest and enjoyment were deduced from the questionnaire compiled by the subjects. The five possibilities given for scoring how much subjects liked the stimuli were converted into a score from 0 to 10, where 0 corresponds to "strongly disliked", 5 corresponds to "neutral" and 10 corresponds to "strongly liked".

Motivation to speak

Motivation to speak was measured as the percentage of time the subjects were speaking or producing any sound while watching at the stimuli. The measure was made given that silence and pauses were expected and deemed appropriate in some of the stimuli. The percentage figure was converted into a continuous 0 - 10 scale.

Significant changes in loudness

The term "significant" meant that the stimuli caused noticeable and consistent changes in loudness which followed the rhythm of the visual stimuli. This was rated with a continuous score between 0 (no changes in loudness) and 10 (highly significant changes which followed the rhythm of the visual stimuli), taking into account the loudness range of the subject. A score of 5 was defined as the minimum acceptable "significant" value.

Significant changes in pitch

As in the case of loudness, a continuous score between 0 and 10 was used to rate the degree of change in pitch following the stimuli, and its consistency. The score takes into account the pitch range of the subject.

Significant changes in vowel quality

Changes in vowel quality were considered with care. When it was clear that the subject was trying to say some word to describe what was going on the screen (for example "high", "low", or "up", "down") this was not considered a significant change in vowel quality, since it was clear that the stimulus was not suggesting directly a particular vowel. Again, a continuous value from 0 to 10 was used.

Production of fricatives

As was expected, some subjects produced fricatives while watching some of the stimuli. During the evaluation of results, this feature was scored as the percentage of time the subject was producing fricatives with respect to the production of any other sound, with 0 meaning 0% and 10 meaning 100% (for example a value of 4 means that when the subject was not silent, they were producing fricatives for 40% of the overall time they were producing a sound). No attempt to characterise different types of fricatives was done. However, it was decided not to include this data in the results, since it is not clear how to take advantage of this information. A significant change in loudness, pitch or vowel quality immediately suggests the use of that stimulus for rehabilitating these speech features. A visual stimulus suggesting the production of fricatives, on the other hand, has a less clear application, due to the more complex issues related to consonant rehabilitation.

Production of "CV" sequences

Some stimuli caused the subject to produce "CV" (consonant-vowel) sequences such as "pa-pa", "ta-ta", etc. As with to the production of fricatives, this feature was scored as the percentage of time the subject was producing CV sequences with respect to the production of any other sound, with 0 meaning 0% and 10 meaning 100% of the time. A note of the actual CV sequences was also taken. However, this data was not included in the results for the same reasons given above.

Overall "interest rating" for the stimuli

The ultimate goal of the evaluation of results of this experiment was to select visual stimuli that were interesting for feedback in voice rehabilitation. It must be noted that higher "significant change" in some of the speech features is not sufficient alone to make a stimulus interesting. It is difficult to change pitch and loudness independently, even for normal-hearing people. Generally a rise in pitch is accompanied by a rise in loudness, and vice-versa. It is therefore interesting to study if some of the

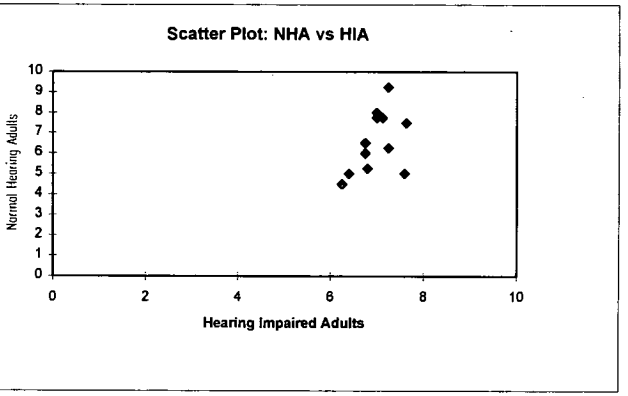
stimuli are able to suggest a change that mostly affects only one of these two features. For this reason a stimulus that was rated 8 both in "loudness" and "pitch" may be less interesting than one rated 6 in pitch only. The "overall interest rating" rates the stimuli taking into account these considerations. However this method for considering a stimulus "interesting" can be arguable, depending on the speech rehabilitation methodology being used. In some cases (generally with children) the use of visual feedback may be preferred although it is less able to make the user control two different speech features but which is nonetheless highly motivating. Another visual feedback may achieve better results in separating two speech features, but it may be less "convincing" or enjoyable. For this reason the "overall interest rating" values are not shown in the results charts, but they are discussed in the evaluation of results instead.

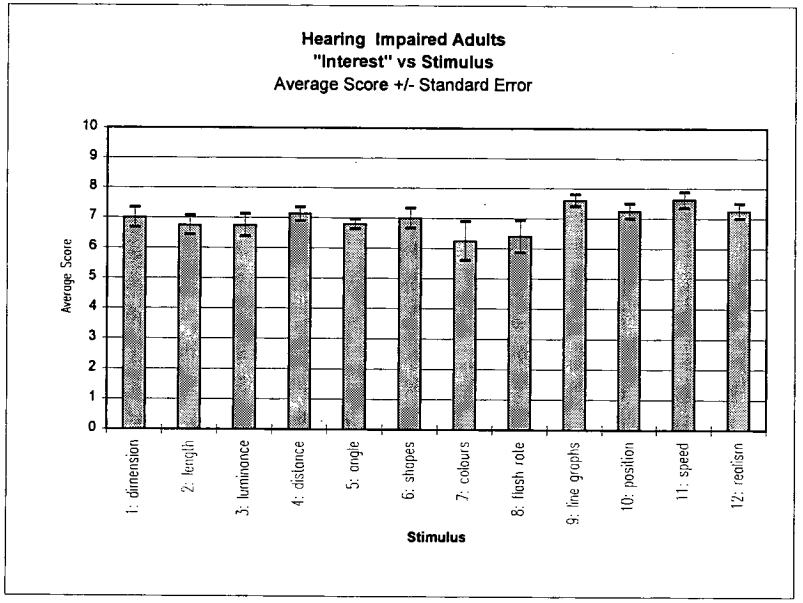
5.2.2 Results

Results for Interest, Motivation to Speak, Significant Changes in Loudness, Significant Changes in Pitch, Significant Changes in Vowel Quality, comparison between Loudness, Pitch and Vowel Quality, are shown in the following Figures (from Figure 5.2 to Figure 5.13) for both groups of hearing-impaired and normal hearing subjects[1]. Table 5.3 to Table 5.9 report the result of a series of t-Tests, as described later in Section 5.2.3.

[1] Abbreviations used in the Figures: HIA: Hearing-Impaired Adults; HIC: Hearing-Impaired Children; NHA: Normal-Hearing Adults; NHC: Normal-Hearing Children.

Interest						
	Hearing Impaired Adults			Normal Hearing Adults		
Stimuli	Av Score	SD	SE	Av Score	SD	SE
1: dimension	7.0	0.9	0.33	7.8	1.8	0.58
2: length	6.8	0.9	0.31	6.0	2.4	0.76
3: luminance	6.8	1.0	0.37	6.5	1.7	0.55
4: distance	7.1	0.6	0.23	7.8	2.2	0.69
5: angle	6.8	0.4	0.16	5.3	2.8	0.87
6: shapes	7.0	0.9	0.33	8.0	2.3	0.73
7: colours	6.3	1.8	0.65	4.5	2.8	0.90
8: flash rate	6.4	1.5	0.54	5.0	1.7	0.53
9: line graphs	7.6	0.5	0.19	5.0	2.9	0.91
10: position	7.3	0.7	0.25	6.3	2.1	0.67
11: speed	7.6	0.7	0.26	7.5	2.4	0.75
12: realism	7.3	0.7	0.25	9.3	1.2	0.38
Average	6.98			6.56		
SE	0.12			0.43		





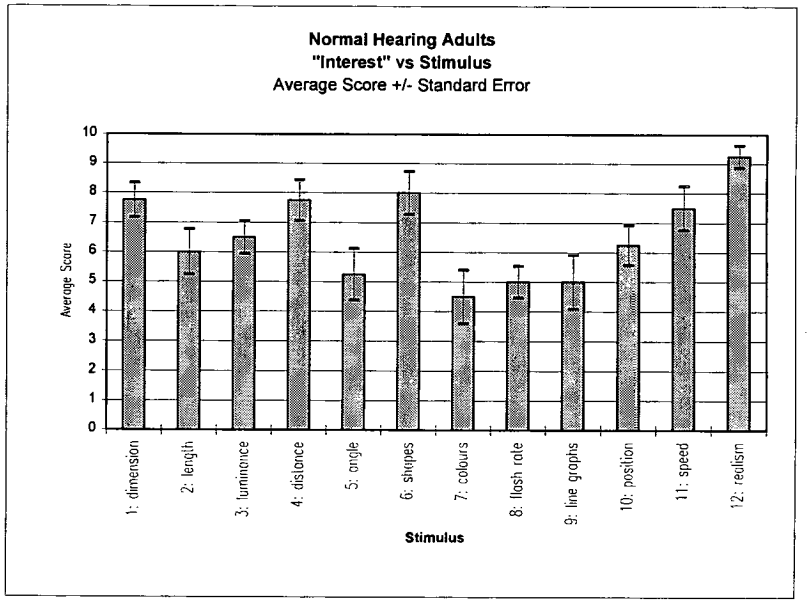
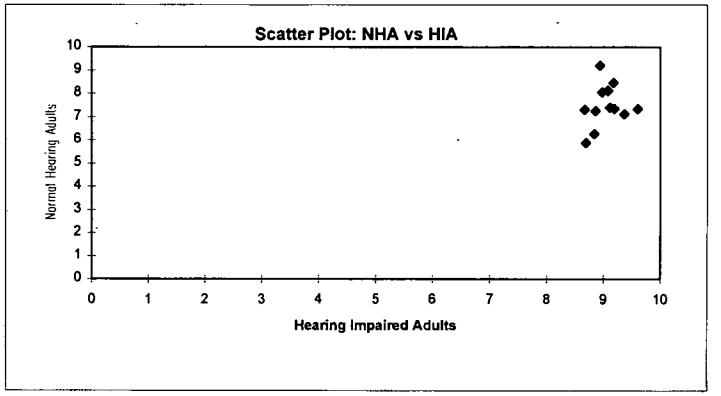
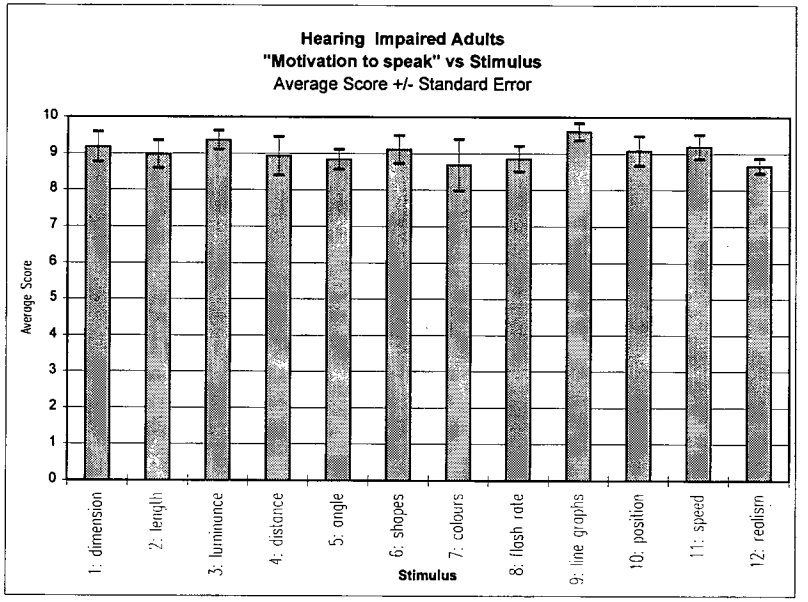


Figure 5.2. Experiment 1: Interest (Adults)

Motivation to speak						
	Hearing Impaired Adults			Normal Hearing Adults		
Stimuli	Av Score	SD	SE	Av Score	SD	SE
1: dimension	9.2	1.2	0.41	8.5	2.0	0.63
2: length	9.0	1.1	0.38	8.1	2.4	0.77
3: luminance	9.4	0.7	0.26	7.1	3.0	0.95
4: distance	8.9	1.5	0.53	9.2	1.6	0.52
5: angle	8.8	0.8	0.27	6.3	2.4	0.77
6: shapes	9.1	1.1	0.38	7.4	3.4	1.09
7: colours	8.7	2.0	0.71	5.9	3.9	1.25
8: flash rate	8.9	1.0	0.35	7.3	3.2	1.00
9: line graphs	9.6	0.7	0.23	7.4	4.0	1.26
10: position	9.1	1.2	0.41	8.1	3.5	1.09
11: speed	9.2	0.9	0.33	7.4	3.8	1.21
12: realism	8.7	0.6	0.20	7.3	2.6	0.81
Average	9.04			7.48		
SE	0.08			0.26		





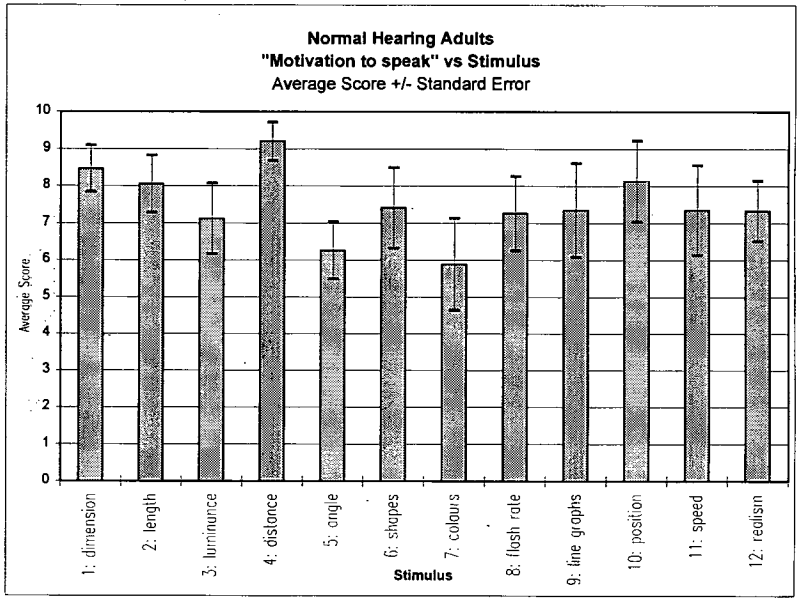
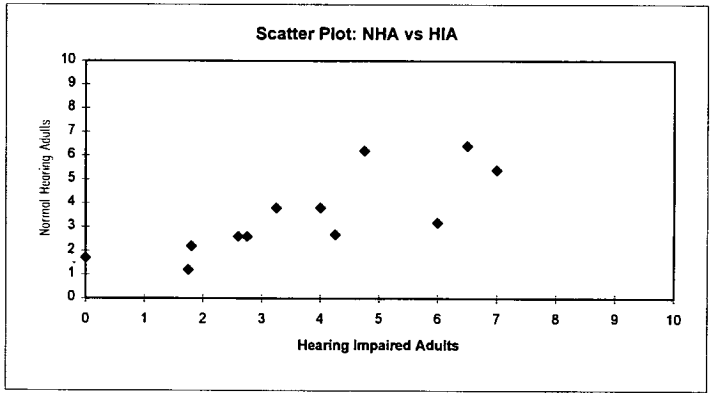
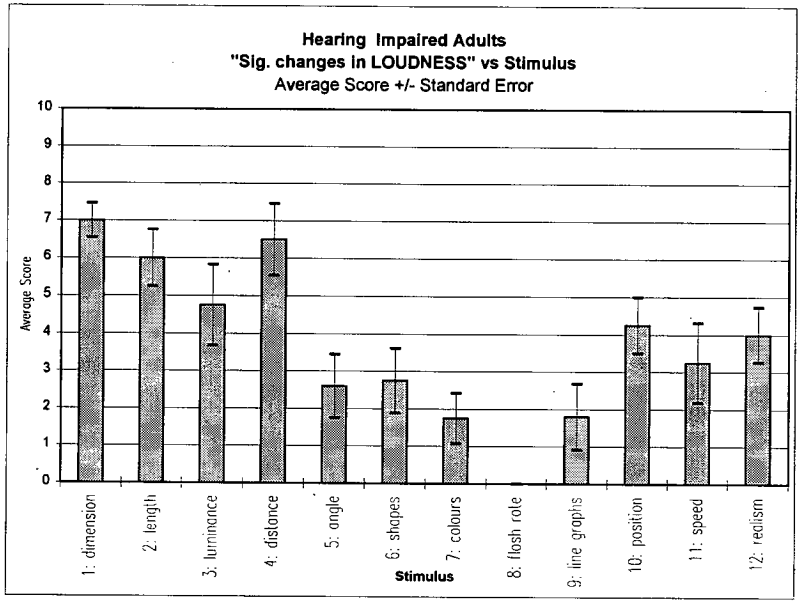


Figure 5.3. Experiment 1: Motivation to Speak (Adults)

Significant Changes in Loudness						
	Hearing Impaired Adults			Normal Hearing Adults		
Stimuli	Av Score	SD	SE	Av Score	SD	SE
1: dimension	7.0	1.3	0.46	5.4	2.2	0.69
2: length	6.0	2.1	0.76	3.2	1.4	0.44
3: luminance	4.8	3.1	1.08	6.2	1.9	0.61
4: distance	6.5	2.7	0.96	6.4	2.8	0.90
5: angle	2.6	2.4	0.85	2.6	1.6	0.52
6: shapes	2.8	2.4	0.86	2.6	1.6	0.52
7: colours	1.8	1.9	0.67	1.2	1.4	0.44
8: flash rate	0.0	0.0	0.00	1.7	2.5	0.79
9: line graphs	1.8	2.5	0.88	2.2	2.2	0.70
10: position	4.3	2.1	0.75	2.7	1.9	0.62
11: speed	3.3	3.0	1.06	3.8	2.3	0.74
12: realism	4.0	2.1	0.73	3.8	0.9	0.29
Average	3.72			3.48		
SE	0.61			0.49		





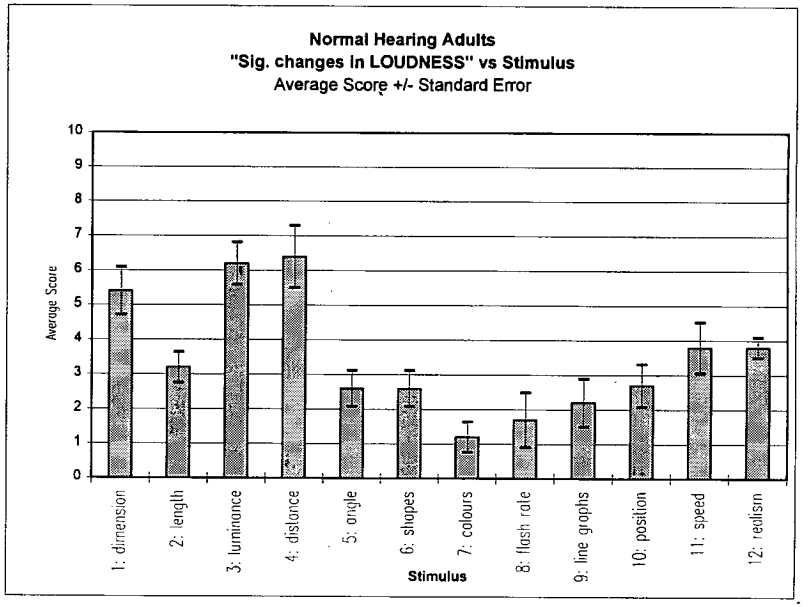
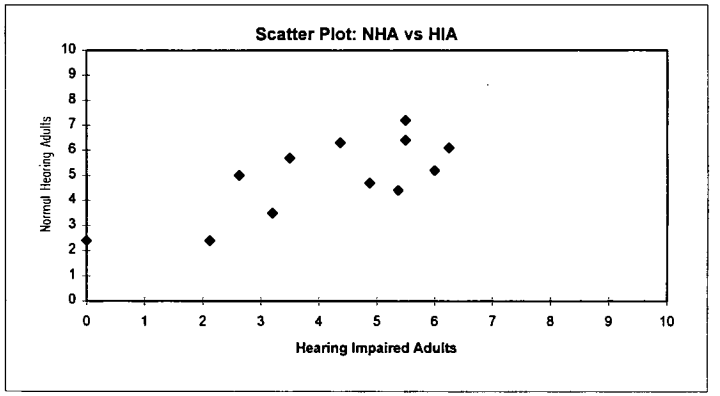
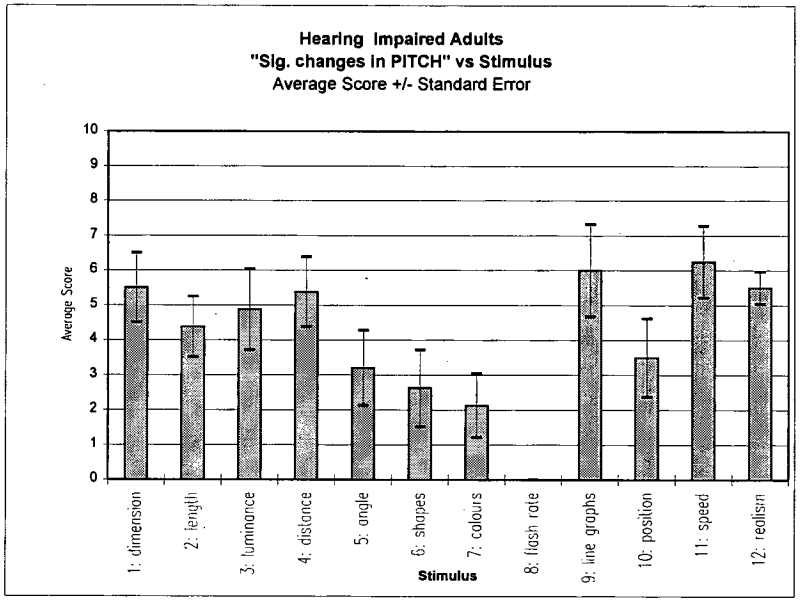


Figure 5.4. Experiment 1: Significant Changes in Loudness (Adults)

Significant Changes in Pitch						
	Hearing Impaired Adults			Normal Hearing Adults		
Stimuli	Av Score	SD	SE	Av Score	SD	SE
1: dimension	5.5	2.8	1.00	7.2	1.6	0.51
2: length	4.4	2.4	0.86	6.3	2.8	0.79
3: luminance	4.9	3.3	1.16	4.7	2.8	0.91
4: distance	5.4	2.8	1.00	4.4	3.5	1.07
5: angle	3.2	3.0	1.07	3.5	2.3	0.64
6: shapes	2.6	3.1	1.10	5.0	2.9	0.80
7: colours	2.1	2.6	0.91	2.4	2.7	0.81
8: flash rate	0.0	0.0	0.00	2.4	2.3	0.93
9: line graphs	6.0	3.7	1.32	5.2	3.4	1.17
10: position	3.5	3.2	1.12	5.7	2.4	0.72
11: speed	6.3	2.9	1.03	6.1	3.5	1.08
12: realism	5.5	1.3	0.46	6.4	1.4	0.43
Average	4.11			4.94		
SE	0.54			0.45		





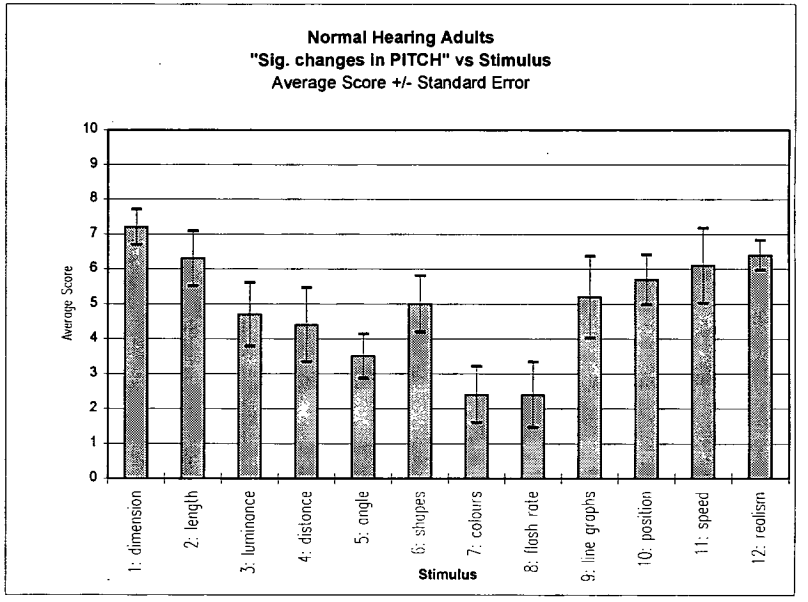
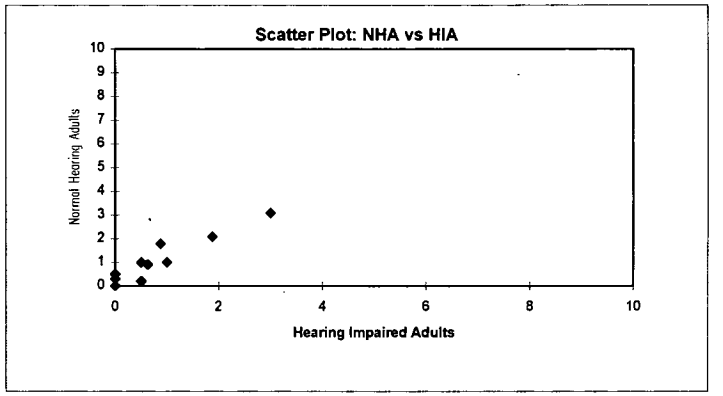
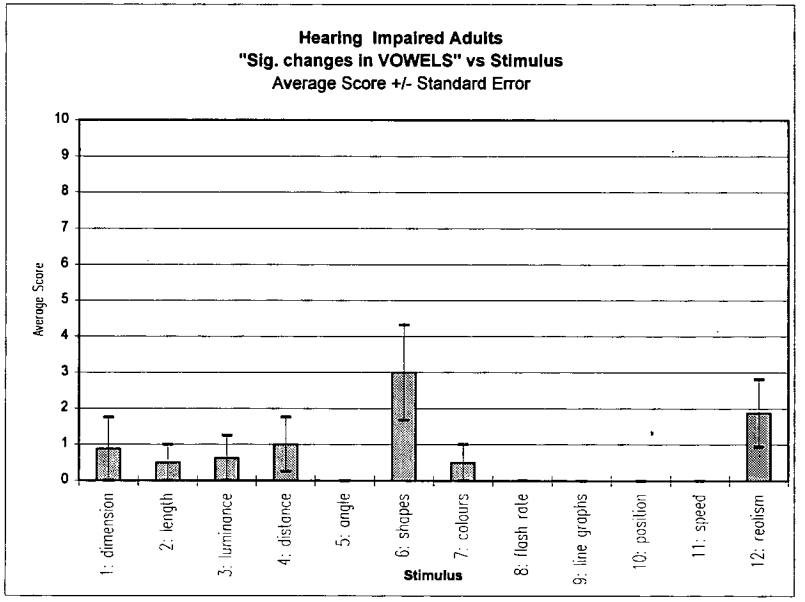


Figure 5.5. Experiment 1: Significant Changes in Pitch (Adults)

Significant Changes in Vowels						
	Hearing Impaired Adults			Normal Hearing Adults		
Stimuli	Av Score	SD	SE	Av Score	SD	SE
1: dimension	0.9	2.5	0.88	1.8	2.6	0.77
2: length	0.5	1.4	0.50	0.2	0.0	0.20
3: luminance	0.6	1.8	0.63	0.9	2.5	0.71
4: distance	1.0	2.1	0.76	1.0	1.8	0.56
5: angle	0.0	0.0	0.00	0.3	1.1	0.30
6: shapes	3.0	3.7	1.32	3.1	3.4	1.03
7: colours	0.5	1.4	0.50	1.0	1.9	0.56
8: flash rate	0.0	0.0	0.00	0.3	1.1	0.30
9: line graphs	0.0	0.0	0.00	0.0	0.0	0.00
10: position	0.0	0.0	0.00	0.5	0.4	0.40
11: speed	0.0	0.0	0.00	0.0	0.0	0.00
12: realism	1.9	2.6	0.93	2.1	2.1	0.81
Average	0.70			0.93		
SE	0.27			0.28		





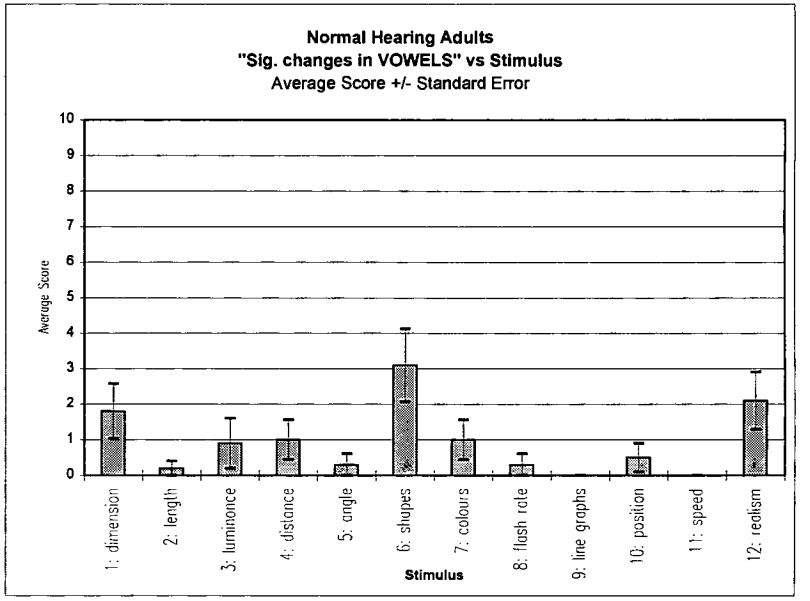
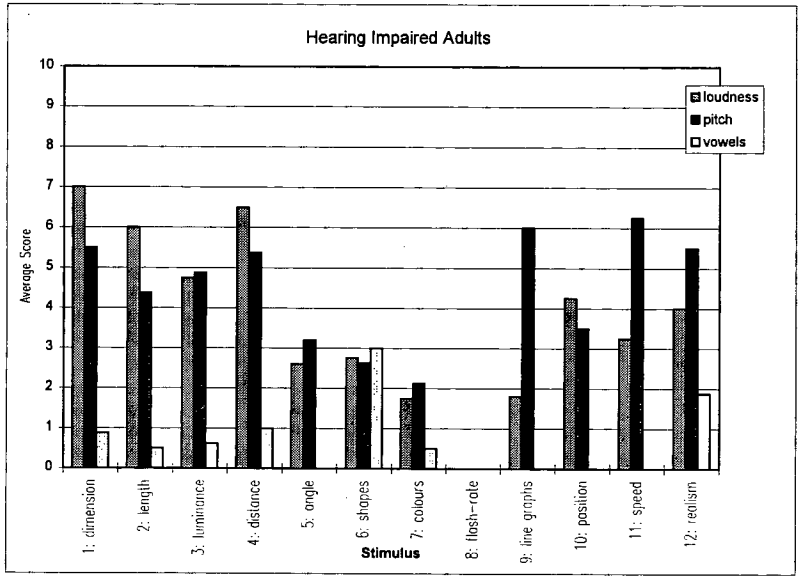


Figure 5.6. Experiment 1: Significant Changes in Vowel Quality (Adults)

Averages						
	Hearing Impaired Adults			Normal Hearing Adults		
Stimuli	Loudness	Pitch	Vowels	Loudness	Pitch	Vowels
1: dimension	7.0	5.5	0.88	5.4	7.2	1.80
2: length	6.0	4.4	0.50	3.2	6.3	0.20
3: luminance	4.8	4.9	0.63	6.2	4.7	0.90
4: distance	6.5	5.4	1.00	6.4	4.4	1.00
5: angle	2.6	3.2	0.00	2.6	3.5	0.30
6: shapes	2.8	2.6	3.00	2.6	5.0	3.10
7: colours	1.8	2.1	0.50	1.2	2.4	1.00
8: flash rate	0.0	0.0	0.00	1.7	2.4	0.30
9: line graphs	1.8	6.0	0.00	2.2	5.2	0.00
10: position	4.3	3.5	0.00	2.7	5.7	0.50
11: speed	3.3	6.3	0.00	3.8	6.1	0.00
12: realism	4.0	5.5	1.88	3.8	6.4	2.10
Average	3.72	4.11	0.70	3.48	4.94	0.93



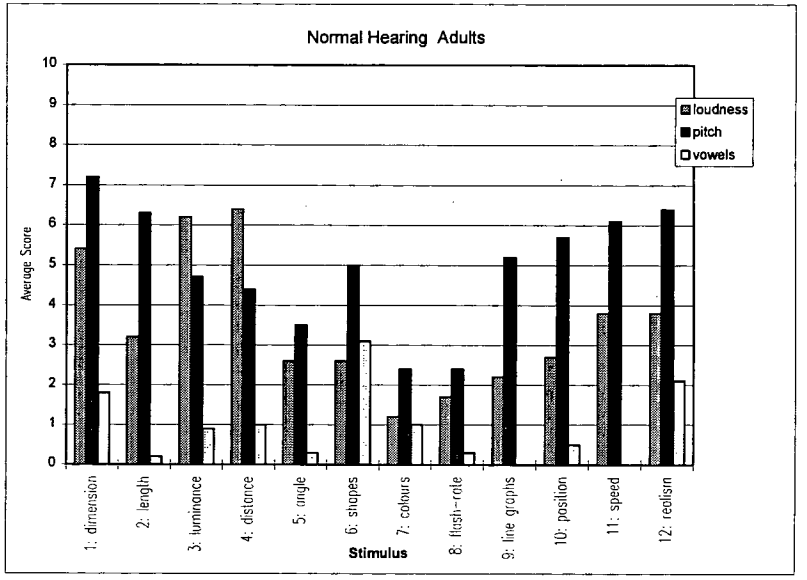
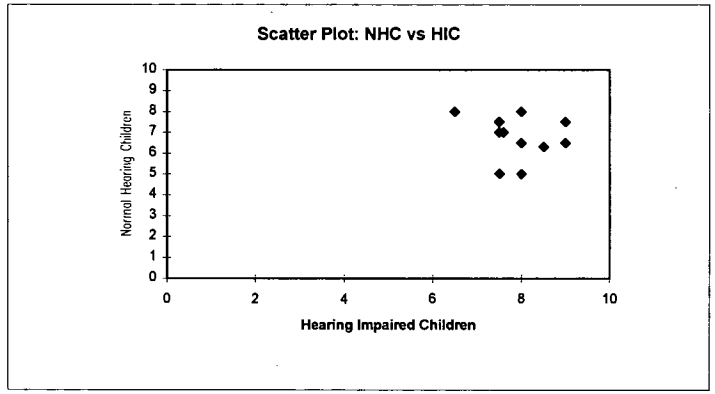
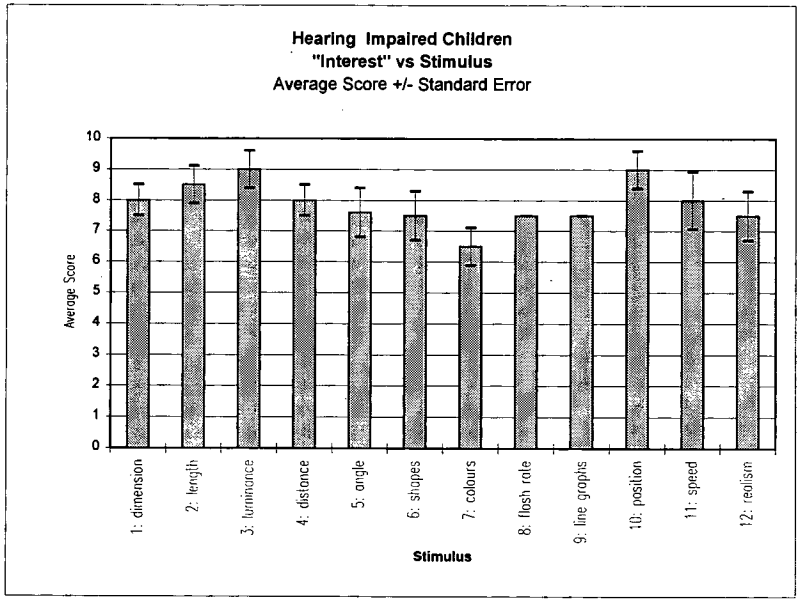


Figure 5.7. Experiment 1: Loudness, Pitch and Vowel Quality averages (Adults)

Interest						
	Hearing Impaired Children			Normal Hearing Children		
Stimuli	Av Score	SD	SE	Av Score	SD	SE
1: dimension	8.0	1.1	0.50	5.0	4.0	1.77
2: length	8.5	1.4	0.61	6.3	4.1	1.85
3: luminance	9.0	1.4	0.61	7.5	1.8	0.79
4: distance	8.0	1.1	0.50	6.5	4.2	1.87
5: angle	7.6	1.8	0.80	7.0	2.1	0.94
6: shapes	7.5	1.8	0.79	7.0	4.1	1.84
7: colours	6.5	1.4	0.61	8.0	3.3	1.46
8: flash rate	7.5	0.0	0.00	7.0	4.5	2.00
9: line graphs	7.5	0.0	0.00	5.0	4.0	1.77
10: position	9.0	1.4	0.61	6.5	4.2	1.87
11: speed	8.0	2.1	0.94	8.0	4.5	2.00
12: realism	7.5	1.8	0.79	7.5	2.5	1.12
Average	7.88			6.78		
SE	0.20			0.29		





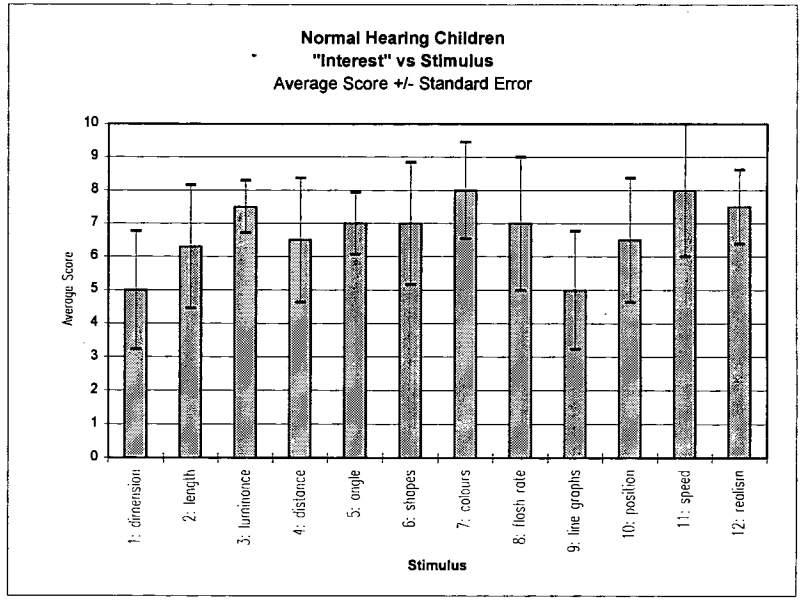
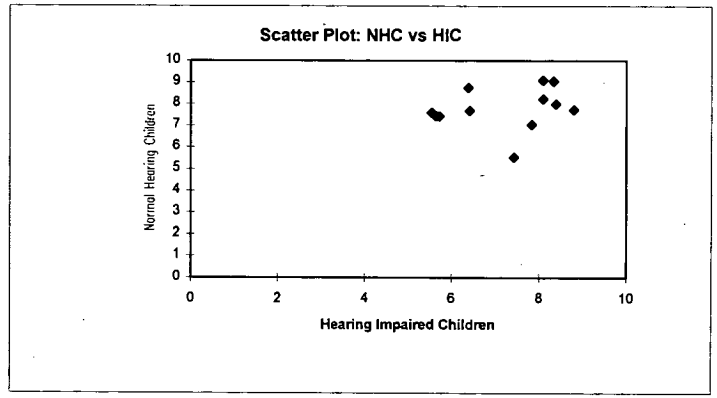
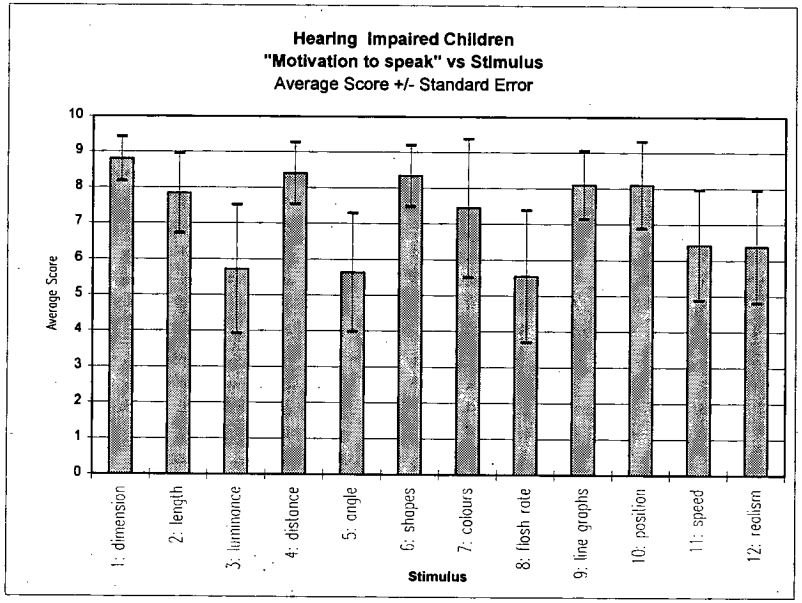


Figure 5.8. Experiment 1: Interest (Children)

Motivation to Speak						
	Hearing Impaired Children			Normal Hearing Children		
Stimuli	Av Score	SD	SE	Av Score	SD	SE
1: dimension	8.8	1.4	0.62	7.7	3.8	1.69
2: length	7.8	2.5	1.11	7.0	2.8	1.25
3: luminance	5.7	4.0	1.80	7.4	2.1	0.95
4: distance	8.4	1.9	0.87	8.0	2.2	1.00
5: angle	5.6	3.7	1.65	7.4	0.4	0.17
6: shapes	8.3	1.9	0.86	9.1	0.8	0.34
7: colours	7.4	4.3	1.94	5.5	5.1	2.26
8: flash rate	5.5	4.1	1.84	7.6	0.7	0.33
9: line graphs	8.1	2.1	0.95	9.1	1.1	0.48
10: position	8.1	2.7	1.21	8.2	2.1	0.94
11: speed	6.4	3.5	1.55	7.7	1.0	0.44
12: realism	6.4	3.5	1.57	8.8	0.4	0.19
Average	7.23			7.80		
SE	0.35			0.28		





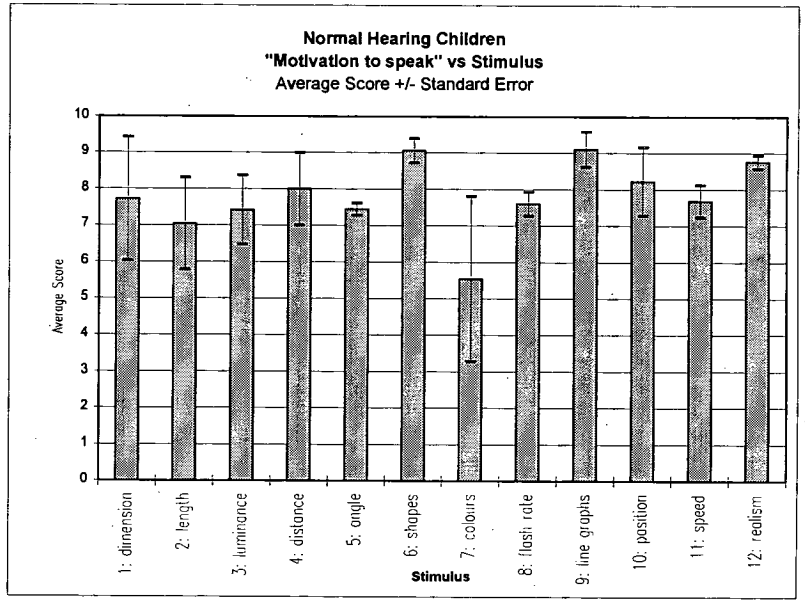
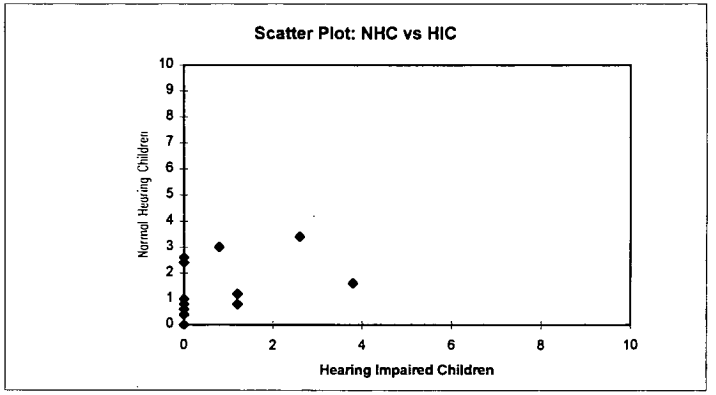
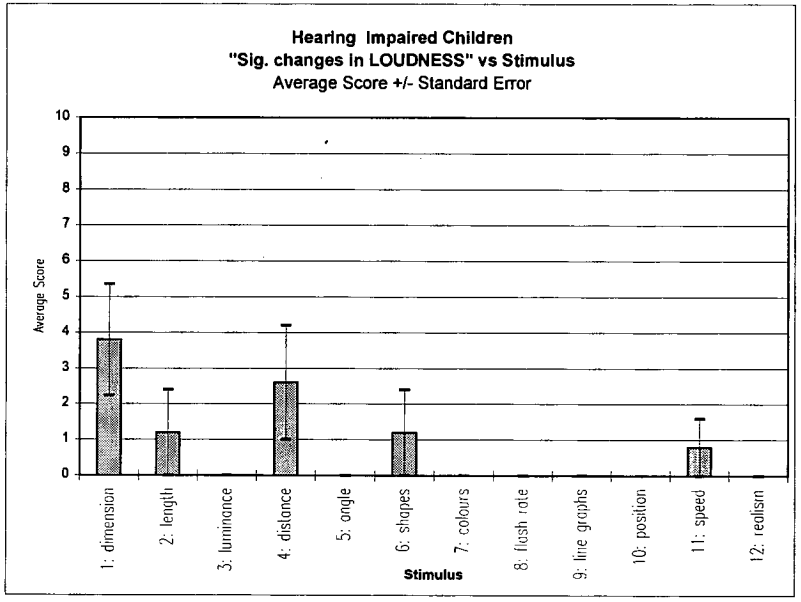


Figure 5.9. Experiment 1: Motivation to Speak (Children)

Significant Changes in Loudness						
	Hearing Impaired Children			Normal Hearing Children		
Stimuli	Av Score	SD	SE	Av Score	SD	SE
1: dimension	3.8	3.5	1.56	1.6	1.5	0.68
2: length	1.2	2.7	1.20	1.2	1.1	0.49
3: luminance	0.0	0.0	0.00	0.0	0.0	0.00
4: distance	2.6	3.6	1.60	3.4	2.5	1.12
5: angle	0.0	0.0	0.00	1.0	1.4	0.63
6: shapes	1.2	2.7	1.20	0.8	1.1	0.49
7: colours	0.0	0.0	0.00	0.6	1.3	0.60
8: flash rate	0.0	0.0	0.00	0.4	0.9	0.40
9: line graphs	0.0	0.0	0.00	2.6	0.5	0.24
10: position	0.0	0.0	0.00	0.8	1.1	0.49
11: speed	0.8	1.8	0.80	3.0	2.6	1.18
12: realism	0.0	0.0	0.00	2.4	1.8	0.81
Average	0.80			1.48		
SE	0.36			0.32		





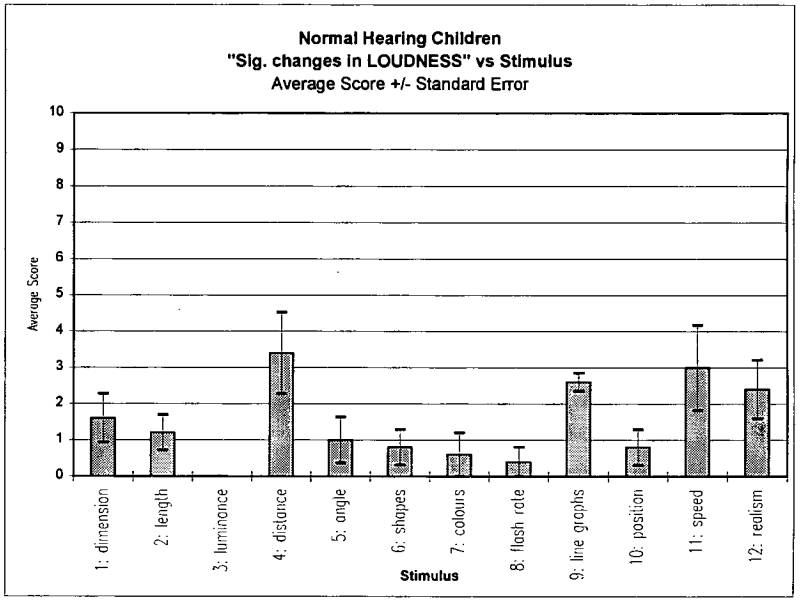
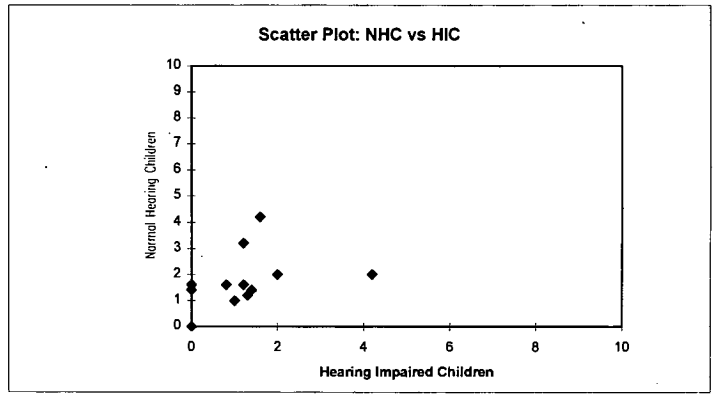
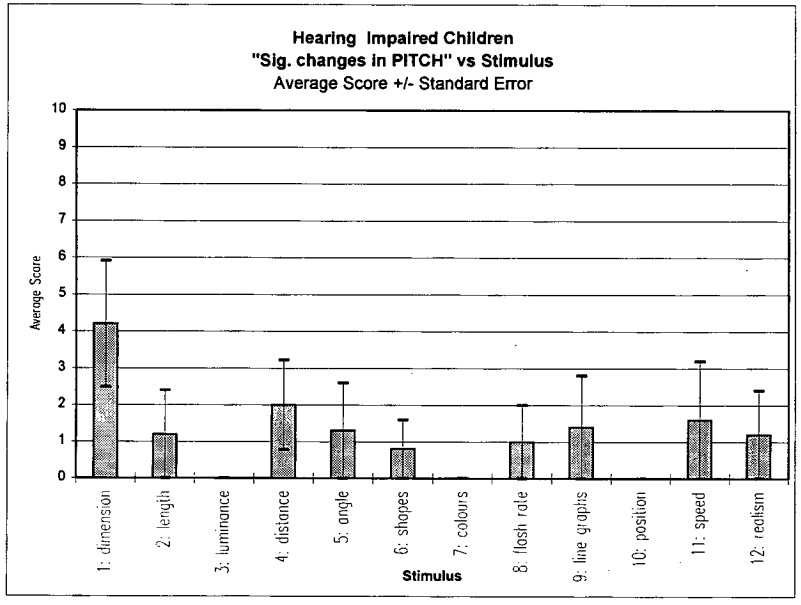


Figure 5.10. Experiment 1: Significant Changes in Loudness (Children)

Significant Changes in Pitch						
	Hearing Impaired Children			Normal Hearing Children		
Stimuli	Av Score	SD	SE	Av Score	SD	SE
1: dimension	4.2	3.8	1.71	2.0	3.5	1.55
2: length	1.2	2.7	1.20	1.6	3.0	1.36
3: luminance	0.0	0.0	0.00	0.0	0.0	0.00
4: distance	2.0	2.7	1.22	2.0	2.4	1.10
5: angle	1.3	2.9	1.30	1.2	1.8	0.80
6: shapes	0.8	1.8	0.80	1.6	1.7	0.75
7: colours	0.0	0.0	0.00	1.4	1.9	0.87
8: flash rate	1.0	2.2	1.00	1.0	1.4	0.63
9: line graphs	1.4	3.1	1.40	1.4	1.3	0.60
10: position	0.0	0.0	0.00	1.6	2.6	1.17
11: speed	1.6	3.6	1.60	4.2	1.6	0.73
12: realism	1.2	2.7	1.20	3.2	2.8	1.24
Average	1.23			1.77		
SE	0.33			0.31		





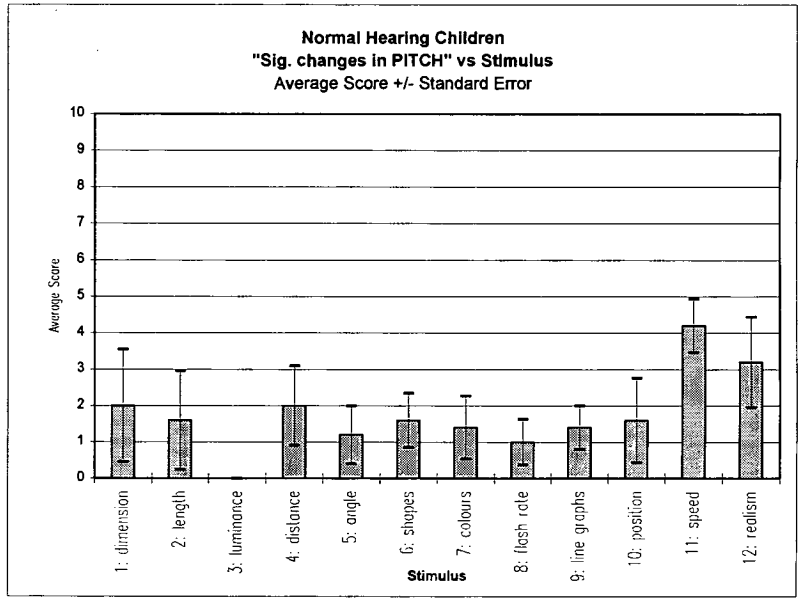
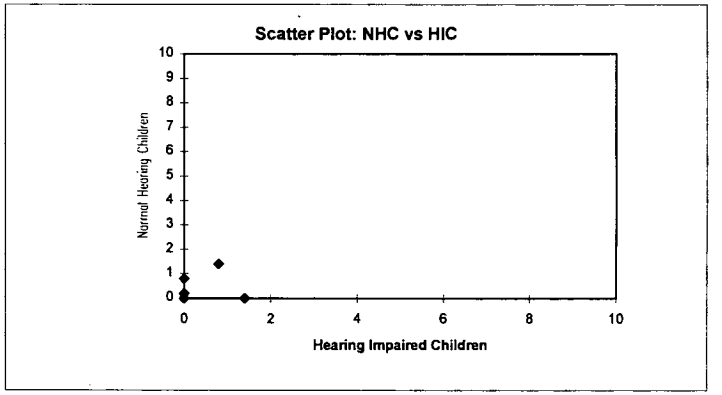
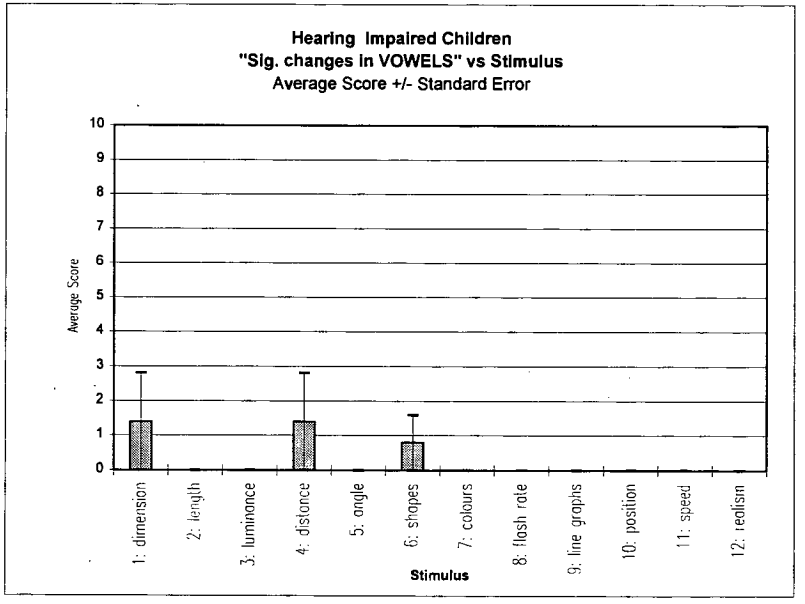


Figure 5.11. Experiment 1: Significant Changes in Pitch (Children)

Significant Changes in Vowel Quality						
	Hearing Impaired Children			Normal Hearing Children		
Stimuli	Av Score	SD	SE	Av Score	SD	SE
1: dimension	1.4	3.1	1.40	0.0	0.0	0.00
2: length	0.0	0.0	0.00	0.2	0.4	0.20
3: luminance	0.0	0.0	0.00	0.0	0.0	0.00
4: distance	1.4	3.1	1.40	0.0	0.0	0.00
5: angle	0.0	0.0	0.00	0.0	0.0	0.00
6: shapes	0.8	1.8	0.80	1.4	3.1	1.40
7: colours	0.0	0.0	0.00	0.8	1.8	0.80
8: flash rate	0.0	0.0	0.00	0.0	0.0	0.00
9: line graphs	0.0	0.0	0.00	0.0	0.0	0.00
10: position	0.0	0.0	0.00	0.0	0.0	0.00
11: speed	0.0	0.0	0.00	0.0	0.0	0.00
12: realism	0.0	0.0	0.00	0.0	0.0	0.00
Average	0.30			0.20		
SE	0.16			0.13		





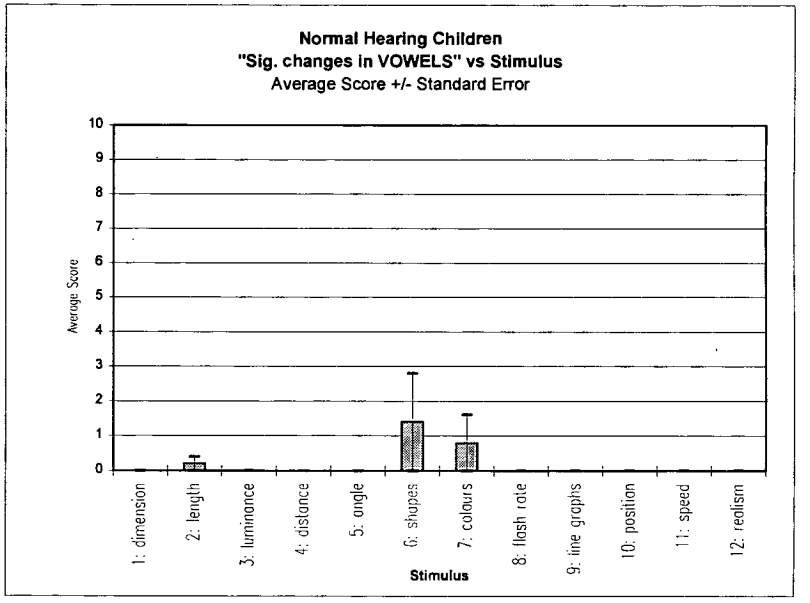
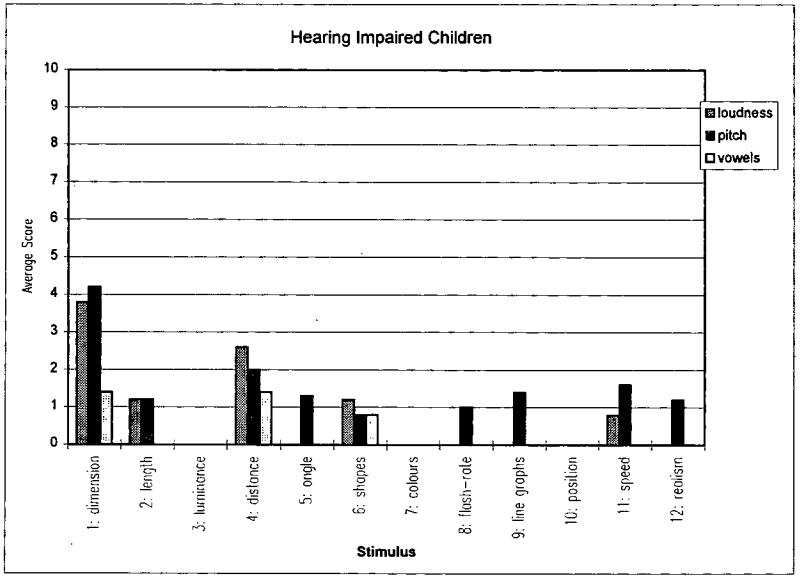


Figure 5.12. Experiment 1: Significant Changes in Vowel Quality (Children)

Averages						
	Hearing Impaired Children			Normal Hearing Children		
Stimuli	**Loudness**	**Pitch**	**Vowels**	**Loudness**	**Pitch**	**Vowels**
1: dimension	3.8	4.2	1.40	1.6	2.0	0.00
2: length	1.2	1.2	0.00	1.2	1.6	0.20
3: luminance	0.0	0.0	0.00	0.0	0.0	0.00
4: distance	2.6	2.0	1.40	3.4	2.0	0.00
5: angle	0.0	1.3	0.00	1.0	1.2	0.00
6: shapes	1.2	0.8	0.80	0.8	1.6	1.40
7: colours	0.0	0.0	0.00	0.6	1.4	0.80
8: flash rate	0.0	1.0	0.00	0.4	1.0	0.00
9: line graphs	0.0	1.4	0.00	2.6	1.4	0.00
10: position	0.0	0.0	0.00	0.8	1.6	0.00
11: speed	0.8	1.6	0.00	3.0	4.2	0.00
12: realism	0.0	1.2	0.00	2.4	3.2	0.00
Average	0.80	1.23	0.30	1.48	1.77	0.20



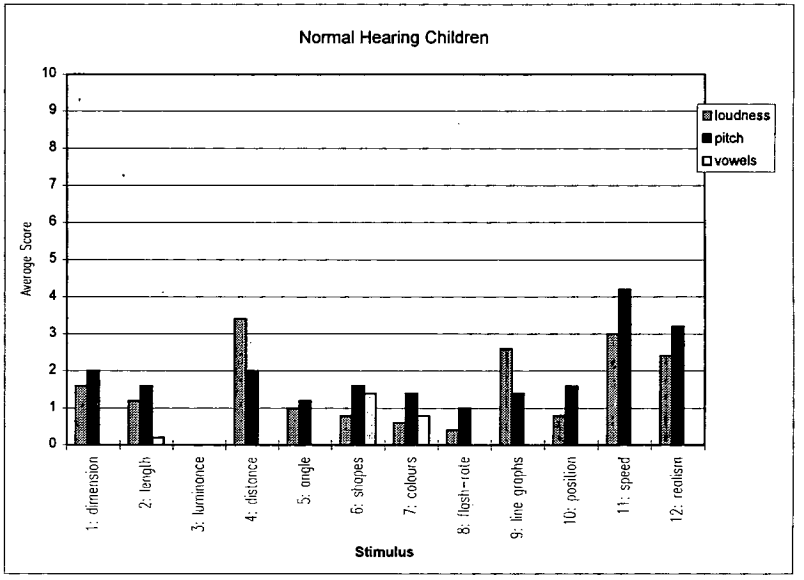


Figure 5.13. Experiment 1: Loudness, Pitch and Vowel Quality averages (Children)

Hearing Impaired Adults - Motivation to speak - t-Test											
	1 dimension	2 length	3 luminance	4 distance	5 angle	6 shapes	7 colours	8 flash rate	9 line graphs	10 position	11 speed
2 length	0.516										
3 luminance	0.432	0.216									
4 distance	0.603	0.944	0.390								
5 angle	0.200	0.112	0.079	0.630							
6 shapes	0.899	0.605	0.576	0.793	0.283						
7 colours	0.557	0.741	0.411	0.617	0.754	0.634					
8 flash rate	0.409	0.371	0.403	0.307	0.972	0.364	0.599				
9 line graphs	0.436	0.461	0.591	0.608	0.110	0.864	0.391	0.246			
10 position	0.833	0.830	0.389	0.808	0.792	0.943	0.653	0.848	0.483		
11 speed	0.958	0.393	0.490	0.534	0.133	0.852	0.526	0.201	0.801	0.790	
12 realism	0.272	0.450	0.065	0.614	0.976	0.250	0.974	0.160	0.160	0.299	0.112

Normal Hearing Adults - Motivation to speak - t-Test											
	1 dimension	2 length	3 luminance	4 distance	5 angle	6 shapes	7 colours	8 flash rate	9 line graphs	10 position	11 speed
2 length	0.247										
3 luminance	0.071	0.102									
4 distance	0.127	0.012	0.017								
5 angle	0.019	0.084	0.399	0.012							
6 shapes	0.087	0.379	0.774	0.071	0.254						
7 colours	0.006	0.021	0.261	0.011	0.710	0.021					
8 flash rate	0.134	0.227	0.868	0.022	0.264	0.853	0.111				
9 line graphs	0.370	0.628	0.897	0.213	0.435	0.962	0.236	0.954			
10 position	0.657	0.856	0.191	0.103	0.186	0.454	0.056	0.282	0.658		
11 speed	0.230	0.289	0.777	0.048	0.380	0.952	0.187	0.908	1.000	0.192	
12 realism	0.009	0.018	0.774	0.002	0.265	0.890	0.056	0.988	0.988	0.160	0.976

Table 5.3. Motivation to speak - p values for two-tailed t-tests comparing different stimuli (Adults)

Hearing Impaired Adults - Significant Changes in Loudness - t-Test											
	1 dimension	2 length	3 luminance	4 distance	5 angle	6 shapes	7 colours	8 flash rate	9 line graphs	10 position	11 speed
2 length	0.200										
3 luminance	0.032	0.129									
4 distance	0.381	0.573	0.076								
5 angle	0.014	0.025	0.018	0.008							
6 shapes	0.001	0.034	0.163	0.008	0.862						
7 colours	0.000	0.004	0.044	0.001	0.208	0.342					
8 flash rate	0.000	0.001	0.000	0.000	0.073	0.077	0.184				
9 line graphs	0.008	0.045	0.040	0.015	0.675	0.178	0.772	0.181			
10 position	0.019	0.076	0.685	0.101	0.160	0.260	0.010	0.022	0.153		
11 speed	0.014	0.138	0.404	0.063	0.895	0.719	0.149	0.208	0.736	0.451	
12 realism	0.010	0.178	0.591	0.104	0.444	0.160	0.083	0.080	0.080	0.847	0.483

Normal Hearing Adults - Significant Changes in Loudness - t-Test											
	1 dimension	2 length	3 luminance	4 distance	5 angle	6 shapes	7 colours	8 flash rate	9 line graphs	10 position	11 speed
2 length	0.024										
3 luminance	0.343	0.004									
4 distance	0.363	0.011	0.823								
5 angle	0.002	0.297	0.000	0.003							
6 shapes	0.010	0.239	0.001	0.000	1.000						
7 colours	0.000	0.008	0.000	0.000	0.016	0.055					
8 flash rate	0.003	0.091	0.000	0.002	0.182	0.287	0.596				
9 line graphs	0.012	0.117	0.001	0.006	0.591	0.522	0.138	0.563			
10 position	0.006	0.551	0.000	0.007	0.868	0.906	0.007	0.299	0.485		
11 speed	0.190	0.168	0.035	0.053	0.174	0.126	0.012	0.081	0.049	0.292	
12 realism	0.074	0.239	0.011	0.036	0.081	0.119	0.000	0.037	0.037	0.093	1.000

Table 5.4. Significant Changes in Loudness - p values for two-tailed t-tests comparing different stimuli (Adults)

Hearing Impaired Adults - Significant Changes in Pitch - t-Test											
	1 dimension	2 length	3 luminance	4 distance	5 angle	6 shapes	7 colours	8 flash rate	9 line graphs	10 position	11 speed
2 length	0.357										
3 luminance	0.537	0.705									
4 distance	0.598	0.401	0.626								
5 angle	0.005	0.521	0.018	0.012							
6 shapes	0.063	0.064	0.135	0.054	0.778						
7 colours	0.012	0.072	0.061	0.012	0.284	0.542					
8 flash rate	0.002	0.030	0.001	0.002	0.078	0.193	0.220				
9 line graphs	0.553	0.530	0.591	0.621	0.025	0.150	0.037	0.023			
10 position	0.015	0.468	0.323	0.008	0.208	0.490	0.196	0.029	0.056		
11 speed	0.483	0.224	0.359	0.468	0.052	0.052	0.003	0.018	1.000	0.073	
12 realism	1.000	0.208	0.460	0.885	0.045	0.018	0.002	1.000	1.000	0.081	0.451

Normal Hearing Adults - Significant Changes in Pitch - t-Test											
	1 dimension	2 length	3 luminance	4 distance	5 angle	6 shapes	7 colours	8 flash rate	9 line graphs	10 position	11 speed
2 length	0.337										
3 luminance	0.033	0.288									
4 distance	0.025	0.041	0.830								
5 angle	0.002	0.000	0.340	0.215							
6 shapes	0.082	0.158	0.785	0.572	0.026						
7 colours	0.002	0.003	0.041	0.115	0.193	0.002					
8 flash rate	0.000	0.007	0.088	0.153	0.392	0.099	1.000				
9 line graphs	0.184	0.382	0.651	0.428	0.094	0.808	0.019	0.115			
10 position	0.081	0.434	0.381	0.134	0.002	0.354	0.005	0.020	0.671		
11 speed	0.396	0.856	0.437	0.311	0.061	0.440	0.028	0.006	0.619	0.786	
12 realism	0.335	0.914	0.124	0.115	0.001	0.094	0.001	0.313	0.313	0.363	0.792

Table 5.5. Significant Changes in Pitch - p values for two-tailed t-tests comparing different stimuli (Adults)

Hearing Impaired Adults - Significant Changes in Vowel Quality - t-Test											
	1 dimension	2 length	3 luminance	4 distance	5 angle	6 shapes	7 colours	8 flash rate	9 line graphs	10 position	11 speed
2 length	0.736										
3 luminance	0.351	0.888									
4 distance	0.685	0.626	0.197								
5 angle	0.374	0.374	0.374	0.374							
6 shapes	0.154	0.153	0.103	0.151	0.195						
7 colours	0.736	1.000	0.888	0.626	#	0.153					
8 flash rate	0.374	0.374	0.374	0.374	#	0.195	#				
9 line graphs	0.374	0.374	0.374	0.374	#	0.195	#	#			
10 position	0.351	0.351	0.351	0.227	#	0.058	0.351	#	#		
11 speed	0.351	0.351	0.351	0.227	#	0.058	0.351	#	#	#	
12 realism	0.240	0.282	0.129	0.213	0.180	0.476	0.282	0.180	0.180	0.085	0.085

Normal Hearing Adults - Significant Changes in Vowel Quality - t-Test											
	1 dimension	2 length	3 luminance	4 distance	5 angle	6 shapes	7 colours	8 flash rate	9 line graphs	10 position	11 speed
2 length	0.176										
3 luminance	0.644	0.385									
4 distance	0.498	0.237	0.889								
5 angle	0.236	0.799	0.479	0.336							
6 shapes	0.153	0.027	0.138	0.045	0.014						
7 colours	0.193	0.416	0.816	0.195	0.548	0.016					
8 flash rate	0.236	0.799	0.195	0.259	1.000	0.038	0.548				
9 line graphs	0.094	0.347	0.237	0.107	0.347	0.013	0.211	0.347			
10 position	0.365	0.195	0.660	0.532	0.719	0.056	0.787	0.719	0.247		
11 speed	0.094	0.347	0.237	0.107	0.347	0.013	0.211	0.347	#	0.247	
12 realism	0.710	0.090	0.498	0.244	0.164	0.202	0.107	0.062	0.062	0.195	0.062

Table 5.6. Significant Changes in Vowel Quality - p values for two-tailed t-tests comparing different stimuli (Adults)

Hearing Impaired Children - Motivation to speak - t-Test											
	1 dimension	2 length	3 luminance	4 distance	5 angle	6 shapes	7 colours	8 flash rate	9 line graphs	10 position	11 speed
2 length	0.877										
3 luminance	0.361	0.227									
4 distance	0.499	0.135	0.171								
5 angle	0.269	0.234	0.859	0.167							
6 shapes	0.517	0.175	0.181	0.374	0.175						
7 colours	0.669	0.851	0.088	0.662	0.077	0.684					
8 flash rate	0.364	0.198	0.660	0.148	0.883	0.159	0.101				
9 line graphs	0.772	0.812	0.055	0.783	0.037	0.825	0.588	0.070			
10 position	0.443	0.409	0.242	0.442	0.250	0.552	0.782	0.207	1.000		
11 speed	0.413	0.057	0.605	0.067	0.617	0.076	0.615	0.506	0.218	0.065	
12 realism	0.269	0.060	0.667	0.075	0.661	0.077	0.634	0.228	0.228	0.075	0.937

Normal Hearing Children - Motivation to speak - t-Test											
	1 dimension	2 length	3 luminance	4 distance	5 angle	6 shapes	7 colours	8 flash rate	9 line graphs	10 position	11 speed
2 length	0.690										
3 luminance	0.872	0.841									
4 distance	0.867	0.022	0.734								
5 angle	0.879	0.790	0.983	0.653							
6 shapes	0.404	0.119	0.214	0.244	0.025						
7 colours	0.256	0.251	0.507	0.145	0.472	0.149					
8 flash rate	0.942	0.587	0.878	0.597	0.750	0.002	0.352				
9 line graphs	0.452	0.227	0.196	0.415	0.032	0.945	0.207	0.092			
10 position	0.599	0.409	0.569	0.856	0.472	0.294	0.228	0.514	0.332		
11 speed	0.976	0.572	0.806	0.718	0.671	0.003	0.328	0.825	0.070	0.388	
12 realism	0.542	0.248	0.181	0.502	0.004	0.423	0.222	0.471	0.471	0.542	0.034

Table 5.7. Motivation to speak -p values for two-tailed t-tests comparing different stimuli (Children)

Hearing Impaired Children - Significant Changes in Loudness - t-Test											
	1 dimension	2 length	3 luminance	4 distance	5 angle	6 shapes	7 colours	8 flash rate	9 line graphs	10 position	11 speed
2 length	0.179										
3 luminance	0.072	0.374									
4 distance	0.458	0.296	0.179								
5 angle	0.072	0.374	#	0.179							
6 shapes	0.179	#	0.374	0.296	0.374						
7 colours	0.072	0.374	#	0.179	#	0.374					
8 flash rate	0.072	0.374	#	0.179	#	0.374	#				
9 line graphs	0.072	0.374	#	0.179	#	0.374	#	#			
10 position	0.072	0.374	#	0.179	#	0.374	#	#	#		
11 speed	0.234	0.815	0.374	0.431	0.374	0.815	0.374	0.374	0.374	0.374	
12 realism	0.072	0.374	#	0.179	#	0.374	#	#	#	#	0.374

Normal Hearing Children - Significant Changes in Loudness - t-Test											
	1 dimension	2 length	3 luminance	4 distance	5 angle	6 shapes	7 colours	8 flash rate	9 line graphs	10 position	11 speed
2 length	0.648										
3 luminance	0.078	0.070									
4 distance	0.286	0.063	0.039								
5 angle	0.426	0.815	0.189	0.099							
6 shapes	0.405	0.374	0.178	0.098	0.854						
7 colours	0.298	0.374	0.374	0.019	0.374	0.838					
8 flash rate	0.109	0.178	0.374	0.070	0.529	0.374	0.815				
9 line graphs	0.230	0.052	0.000	0.477	0.078	0.037	0.022	0.004			
10 position	0.242	0.621	0.178	0.098	0.374	1.000	0.704	0.621	0.037		
11 speed	0.404	0.181	0.064	0.648	0.089	0.198	0.033	0.152	0.757	0.097	
12 realism	0.338	0.178	0.042	0.394	0.025	0.195	0.021	0.828	0.828	0.035	0.529

Table 5.8. Significant Changes in Loudness - p values for two-tailed t-tests comparing different stimuli (Children)

Hearing Impaired Children - Significant Changes in Pitch - t-Test											
	1 dimension	2 length	3 luminance	4 distance	5 angle	6 shapes	7 colours	8 flash rate	9 line graphs	10 position	11 speed
2 length	0.142										
3 luminance	0.070	0.374									
4 distance	0.161	0.495	0.178								
5 angle	0.345	0.962	0.374	0.758							
6 shapes	0.096	0.374	0.374	0.284	0.782						
7 colours	0.070	0.374	#	0.178	0.374	0.374					
8 flash rate	0.263	0.914	0.374	0.621	0.878	0.895	0.374				
9 line graphs	0.178	0.927	0.374	0.799	0.965	0.753	0.374	0.845			
10 position	0.070	0.374	#	0.178	0.374	0.374	#	0.374	0.374		
11 speed	0.432	0.866	0.374	0.875	0.374	0.704	0.374	0.788	0.937	0.374	
12 realism	0.142	1.000	0.374	0.714	0.962	0.815	0.374	0.374	0.374	0.374	0.866

Normal Hearing Children - Significant Changes in Pitch - t-Test											
	1 dimension	2 length	3 luminance	4 distance	5 angle	6 shapes	7 colours	8 flash rate	9 line graphs	10 position	11 speed
2 length	0.859										
3 luminance	0.266	0.306									
4 distance	1.000	0.859	0.142								
5 angle	0.374	0.794	0.208	0.621							
6 shapes	0.704	1.000	0.099	0.799	0.374						
7 colours	0.634	0.847	0.184	0.736	0.704	0.778					
8 flash rate	0.446	0.749	0.189	0.519	0.778	0.468	0.704				
9 line graphs	0.704	0.893	0.080	0.529	0.815	0.838	1.000	0.688			
10 position	0.374	1.000	0.242	0.828	0.374	1.000	0.815	0.529	0.866		
11 speed	0.207	0.144	0.005	0.086	0.023	0.065	0.031	0.030	0.000	0.081	
12 realism	0.235	0.412	0.061	0.595	0.047	0.160	0.121	0.266	0.266	0.078	0.508

Table 5.9. Significant Changes in Pitch - p values for two-tailed t-tests comparing different stimuli (Children)

5.2.3 Evaluation of results

Hearing-impaired adults

As most speech therapists suggest, it is difficult to try to characterise an average behaviour in hearing-impaired subjects: they are all different cases. To measure the reliability of the differences in response observed between different types of stimuli, t-tests were performed for all pairs of stimulus types. The results are shown in Tables 5.3 to 5.6. In view of the large number of tests performed (66 for each subject group and dependent variables), it seems appropriate to apply a fairly stringent significance threshold to the results, so as to reduce the number of spurious conclusions likely to be drown. The comparisons mentioned in the discussion below are those with p values below 0.01.

This evaluation of results focuses on *significant changes in loudness, pitch* and *vowel quality*, since these are the aspects most interesting for the scope of this thesis. However, data on *interest* and *motivation to speak* are useful information in the decision of which visual stimuli can be suitable as a visual feedback method, and are reported in Figure 5.2 and Figure 5.3.

With respect of *significant changes in loudness* with hearing-impaired adult subjects (see Figure 5.4), the series of t-Tests reported in Table 5.4 shows that the average value of the stimulus *dimension* can be reliably compared with the average values of the stimuli *shapes, colours, flash rate, line graphs* and *realism*; the value of the stimulus *length* can be reliably compared with the values of the stimuli *colours* and *flash rate*; *luminance* with *flash rate*; *distance* with *angle, shapes, colours, flash rate*; *colours* with *position.* However, these comparisons are not sufficient for defining which is the stimulus with the best reliable average value. It was decided to consider all the four stimuli having the highest average values (the mean minus one standard error for each of these is still above the average score of all other stimuli). In decreasing order of average score they are *dimension, distance, length,* and *luminance.* Stimuli using *dimension* are already used in speech rehabilitation systems, and this experiment confirms that hearing-impaired adults significantly modulate their loudness following the changes of dimensions of an object shown on the screen. It is difficult to evaluate how much of this is due to the intuitiveness of the stimuli, and how much of this is due to the experience of the subjects in the speech rehabilitation sessions they attended, where the therapist may have used the same metaphor size-loudness to teach the subjects how to control their speech (only two adult subjects had made some use of a computer based speech rehabilitation system). In any case, either because visual feedback is intuitive, or because it was easy to learn and retain, it is valid and worth considering. The *dimension* stimuli caused a variation in pitch, as well, (as shown in Figure 5.5 and Figure 5.7) but not so large as the variation in loudness. It has to be remembered here that the aim is to find a visual stimulus able to be linked with only one speech feature. A similar average value was given by the

novel *distance* stimuli. A first result from this analysis is therefore that the dimension *distance* is a good candidate as a feedback for loudness. The *length* stimuli gave results which were just slightly lower, but with the advantage that the pitch was less affected. *Luminance* has a lower average score in comparison with *dimension, distance* and *length,* and it caused changes in pitch of the same extent. However the subjects behaved quite differently. This means that this modality can work with some people better than with others.

For *significant changes in pitch,* the series of t-Tests (reported in Table 5.5) shows reliable comparisons of the stimulus *dimension* with *angle* and *flash rate; luminance* with *flash rate; distance* with *flash rate* and *position*; *colours* with *speed* and *realism.* Again this is not sufficient for deciding which is the stimulus which reliably produced the most significant changes. Among the four stimuli having the highest average score, *speed,* and *line graph* were considered because of their highest average values, and *realism* was considered more interesting than *dimension* since it has a lower standard error. Furthermore, *dimension* showed the highest average score regarding *significant changes in loudness,* and was therefore already considered for such a speech feature. The *line graph* modality is already used in speech therapy. The novel *speed* stimulus, is definitely worth trying as a visual feedback. The *realistic* stimuli, also are worth investigating.

For *significant changes in vowels*, the series of t-Tests (see Table 5.16) does not indicate any reliable comparison. Furthermore, the average scores in Figure 5.6 show a very low value, suggesting that no stimuli were able to convey to the subject the idea of a particular vowel. An attempt to accomplish this goal, (but exhibiting a low average value) was made by the *shapes* stimuli, that in one of its variations showed an oval shape resembling a pair of lips changing from a horizontal shape to a vertical one. This suggested the production of vowels to some subjects but, without information about the internal position of the tongue, it was impossible to differentiate between back and front vowels. Accordingly, subjects used either /a/ and /ɑ/ when the shape was horizontal, and either /i/ and /u/ when the shape was vertical.

In the evaluation of results, data were also collected on production of fricatives. As explained in Section 5.2.1.7 it was not clear how to use these data in order to build an effective visual feedback, (no attempt to characterise different types of fricatives was done) so it was decided not to include them in the charts. In any case it may be interesting to know that some of the stimuli prompted the subjects to produce a "shhh" sound, especially the *angle* stimulus and some variations of the *speed* stimuli and the *realistic* stimuli.

Note that each visual stimulus consisted of one or more variations of the same type (as shown in Table 5.2). The results from the different variations were averaged in the final results for each

stimulus. An exception to this rule was the *speed* stimuli, where only one of the three variations was considered in the final results (since the other two variations gave very poor responses).

In conclusion, the following visual stimuli resulted in being considered interesting as candidates for implementation in visual feedback for particular speech features for hearing impaired adults:

Dimension, linked with loudness

Distance, linked with loudness

Length, linked with loudness

Luminance, linked with loudness (with some subjects)

Line graphs, linked with pitch

Speed, linked with pitch

Realistic stimuli, linked with pitch.

Hearing-impaired children

Children gave more scattered results than those given by adults. None of the stimuli gave average scores above the value of 4.2 for *significant changes in loudness, pitch and vowel quality* (see Figure 5.10, 5.11 and 5.12) and the series of t-Tests (see Table 5.18 and 5.19) gave no information about reliable comparisons of average values[1]. Children have generally more difficulty in coping with experiments like this, which may soon become boring for them. They tended to follow the stimuli with their voice for *distance*, *dimension* and *length* stimuli, but the effect of these stimuli on their voice was not consistent. Furthermore they occasionally produced sounds without paying too much attention to the screen (as was noted from the reflection in the mirror in the videotapes). For this reason children's responses for the proposed visual stimuli are less clear. The solution in this case may be to give variety. A considerable number of different visual feedback types for each speech feature will enable the speech therapist to experiment with them, and find those which work for each case.

Comparison between hearing-impaired subjects and normal-hearing subjects

In order to compare the results from hearing-impaired subjects and normal-hearing subjects, the following procedure was adopted. For each subject, the samples for each group of stimuli were

[1] The t-Test regarding significant changes in Vowels for Children is not included since the high number of null samples caused the calculation of the variance to fail.

normalised by subtracting the average value of the other groups of stimuli, and then dividing the result by the standard deviation of the other groups of stimuli. A t-Test was then used on normalised data. The results of the series of t-Tests gave a value of the confidence of the comparative data shown in Figure 5.7 (adults) and Figure 5.13 (children). In fact, most values (with the exception of the stimuli *length* and *luminance* for Significant Changes in Loudness) are below the confidence level of 95%, meaning that it is not possible take any conclusion on the difference in response from the group of hearing-impaired subjects and normal-hearing subjects. Figure 5.2 to 5.13 and Table 5.3 to 5.9 report data on both hearing-impaired and normal hearing subjects, together with scatterplots between these two groups.

5.2.4 Conclusion

Experiment 1 conducted with hearing-impaired adult subjects gave useful information about intuitive connections between visual stimuli and speech features. Some visual stimuli which are already used as a feedback method in speech rehabilitation systems were confirmed as interesting (*dimension* and *length* for loudness, and *line graphs* for pitch), and novel ones appeared worth investigating (*distance* and *luminance* for loudness, *speed* and *realistic stimuli* for pitch). Hearing-impaired children unfortunately did not give reliable information, and the scattered data of the results confirms that probably a variety of different visual stimuli can offer an acceptable approach. It is therapists' goal to find the appropriate one that works in each single case.

The data from the experiments conducted with normal-hearing adult subjects, reported in this Chapter, although they may give useful information about normal hearing subject's responses to the proposed visual stimuli, could not be reliably used for comparison with hearing-impaired people data. A larger number of subjects is required if a comparative study has to be produced.

Normal-hearing children were the most difficult category to evaluate. They tended to be not very collaborative, and it seemed very difficult to conduct a controlled experiment with them. This was not surprising, however, since the simplicity of some of the stimuli was of little attraction to them, causing boredom and immediate refusal[1]. This did not happen with hearing-impaired children, who seemed more tolerant and disciplined.

[1] As for hearing-impaired children, for normal-hearing children none of the stimuli gave reliable results. The behaviour of the children who attended the experiment was extremely varied, and this affected the results. For example, one child asked how many stimuli were still missing, since he thought there were too many left and he was impatient to finish the session and go and play with other children. He decided from that moment that he did not like any of the stimuli, until one of them, particularly colourful, made him reverse its decision. A little girl said that her father had a computer, and she knew already the things that happened in the computer. She added that some of the stimuli were "tricky", and refused to produce any noise for those ones.

5.2.5 Criticism of experimental design methodology

Investigative, objective measures

Ratings even by expert phoneticians over several sessions, and involving several aspects, must include some variability, subjectivity or inconsistency. One solution is calibration with a control set of data to establish reliability of ratings. Better still would be to find an objective method of rating the subjects' vocalisations, e.g. using a pitch tracker and high quality recordings of all subject's utterances. This method would have the advantage of providing reliable numerical data, which would be suitable for carrying out parametric statistical analysis. Another advantage of high quality recordings is that signal amplitude reflects loudness, and again this could be correlated with appropriate visual stimuli frequency. In addition, changes in vowel quality could be objectively measured with a formant tracker which would relate changes in F_1, F_2, F_3 to the relevant visual stimuli. However, the relationship between perceived vowel quality and formant values is less obvious and less direct than between pitch and F_0 or loudness and dB SPL.

Although acoustical measures could be more reliable than using expert assessors, there are several problems to be considered. Acoustical measures may require carefully microphone set-up, involving calibration. Pitch trackers are not normally used in these applications and may require considerable customisation for this particular task. For example, the pitch range of hearing impaired subjects is more broadly defined than with normal hearing people (see Section 2.4.2), and typical phonation includes a wider variety of voice qualities, which adversely affects the performance of pitch and formant trackers. In addition, reliable formant trackers were not available when the experiment was conducted (see Section 6.3.2.2). In the light of these facts, it is perhaps currently more appropriate to rely on perceptual ratings of vowel quality, such as phoneticians are trained to make, than on acoustic measures.

Small number of subjects

The number of subjects (especially children) needs to be increased. However it is difficult to estimate the minimum number of subjects required for reliable data. Section 5.2.3 shows that the speech responses of normal hearing subjects in Experiment 1 were not a good index of responses given by hearing-impaired subjects, however this finding may not be repeated in larger trials. Perhaps normal hearing subjects could be used in future trials, the validity of this would have to be established by a comparison of normal hearing and hearing-impaired subjects' responses using larger samples in both groups.

5.3 Experiment 2

5.3.1 Methods and Procedures

The goal of the experiment was to compare subject's preferences to two different methods for displaying a graphical stimulus. The first method makes use of a traditional computer monitor, while the second make use of 3D headset.

5.3.1.1 Selection of the stimulus

The stimulus chosen for the experiment was a realistic animation of a roller-coaster, seen from the point of view of the rider (see Figure 5.14). The stimulus was selected because of its suggestive combination of 3D scenery, changes of speed and height, and suitability for both children and adults.

Figure 5.14. The rollercoaster stimulus

5.3.1.2 Implementation of the stimulus

The stimulus relies on more complex graphical methodologies than the stimuli used in Experiment 1. A software 3D rendering engine (Renderware™) running on a fast PC (a Pentium 90 MHz) was used to generate real-time images, at a frame-rate depending on the size of the actual image. With a size of about half of the total screen area, the frame rate was around 15 frames/sec, adequate for the purpose of the experiment since it is fast enough to give a realistic impression of movement. (Bryson, 1995).

The 3D animation[1] was realised in two versions. The first one used 3D techniques and was displayed on a normal computer screen (simulated 3D). The second one added true stereo-vision, since two moving images were generated at the same time and sent to the two 1" colour LCD screens of a 3D headset (see Figure 5.15). In this case the rendering engine generated two images from two different points of view (at the eyes' distance), and interleaved the horizontal lines that composed the two images: odd lines for the left image and even lines for the right image. The 3D headset took the single video signal from the computer, and separated the two images by sending the odd and even lines to the two separate LCD displays. To maintain the correct geometry, each line was doubled at each display.

Figure 5.15. A subject wearing the 3D headset used in the experiment

5.3.1.3 Subjects

As for Experiment 1, the subjects belonged to two categories: hearing-impaired people and normal-hearing people. The hearing-impaired subjects were ten deaf people of both sex, from which five were children and five were adults. They were a subset of the subjects of the Experiment 1. The normal-hearing subjects, of both sex, were five adults aged over twenty, and five children, aged between four and nine, all with no knowledge of speech rehabilitation techniques.

[1] The animation was taken from a demonstration programme included in the package of the rendering engine, and modified in order to make it more suitable for the experiment's purpose (removing of brand names, change of shape of the track, change of some of the bitmaps in the background) to simplify the images.

5.3.1.4 Structure of individual trials

This experiment was run after Experiment 1 in order to avoid introducing bias in the subject's score of the stimuli shown in Experiment 1, from the experience with a considerably more interesting and probably much more enjoyable graphic. In the first part of the experiment (simulated 3D) the stimulus was presented on the computer screen. The subjects were not forced to sit at a fixed distance from the screen, and they were left free to move as they preferred (as is likely in a real therapy session). In the second part of the experiment the subject had to wear the 3D headset[1]. Each of the two parts were shown to the subject for two minutes. Again the subject was asked to try to accompany the images with their voice producing any sound or noise they considered appropriate. The active participation of the subject with their voice was chosen for two reasons: 1) it was easier to assess the level of participation and interest of the subject in the visual stimuli; 2) it was interesting to assess the capability of the stimuli to suggest to the subjects variations in some of the speech features.

5.3.1.5 Interviews

At the end of both sessions the subjects were asked to rate on a form how much they liked the different stimuli, and why. As with Experiment 1 the form had five scores (from 1 to 5) for each stimuli, named "strongly disliked", "disliked", "neutral", "liked", "strongly liked", with accompanying pictures showing "unhappy faces" or "smiling faces" to make things clearer and more attractive for children. There was a space for comments as well.

5.3.2 Results

Hearing-impaired subjects

This set of stimuli was generally very well accepted by all of the hearing-impaired subjects (see Figure 5.16). While adults had no preferences for one method or the other (only one subject preferred the 3D stimuli over the 2D), invariably all the children gave the same answer, that is they liked the 2D stimuli (score 4) and they strongly liked the 3D stimuli (score 5). The only subject, an adult, who gave a "neutral" preference (score 3), motivated her decision with the fact that the stimuli "made her feel sick".

[1] The headset used was the "i-glasses" Virtual IO™. This headset has the following characteristics: horizontal field of view (100% overlap) 30'; horizontal resolution: 263 pixels; vertical resolution: 230 pixels; angular resolution: 6.84 arc minutes.

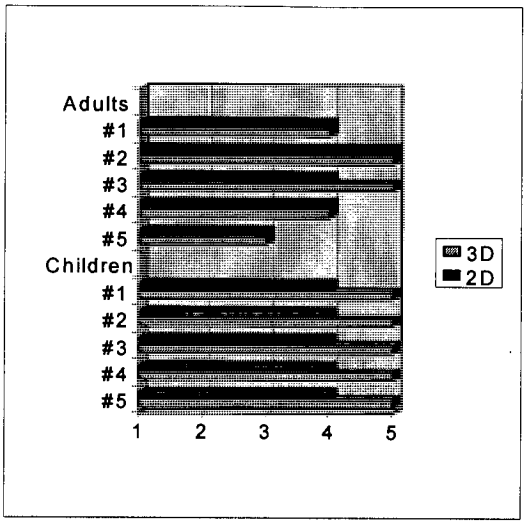


Figure 5.16. Hearing-impaired subject's ratings for Experiment 2

Normal-hearing subjects

Normal-hearing subjects had more various opinions about the stimuli used in Experiment 2 (see Figure 5.17). Three subjects did not like the 3D stimuli (they said it made them feel sick) and gave neutral or negative responses. All the other seven liked or strongly liked both versions, with the exception of an adult and one child who gave neutral responses for the 2D stimuli.

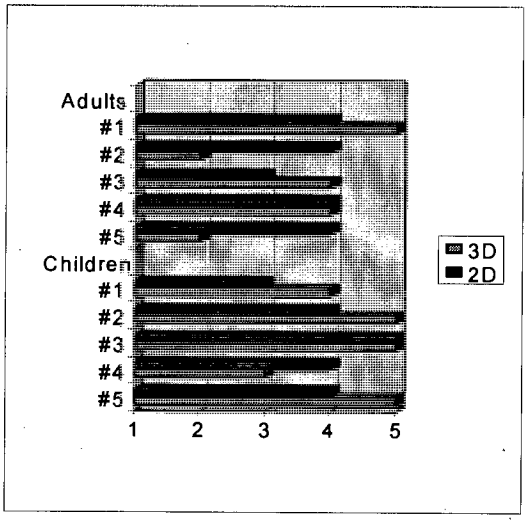


Figure 5.17. Normal-hearing subject's ratings for Experiment 2

5.3.3 Conclusion

Table 5.10 shows the average, standard deviation (SD) and standard error (SE) values for the Experiment 2. It shows that as an average both modalities were similarly accepted. In the case of hearing impaired children, the 3D stimuli gave the highest possible average, indicating that this technology is worth investigating in the field of hearing impaired speech rehabilitation for children, although the still high price of 3D headset may limit their use. However, video-game brands (such as Nintendo and Sega) are marketing a cheap version of a 3D headset. Testing the suitability of such devices for speech therapy applications goes beyond the scope of this thesis, however it appears that this technology will become more common in future.

Criticism of experimental design methodology

A possibly better way to conduct this experiment would be to ask subjects make an explicit preference towards one of the two methods, and compare preference with attitude scores.

	2D			3D		
Subject Group	Average	SD	SE	Average	SD	SE
Hearing-Imp. Adults	4.0	0.7	0.32	4.2	0.8	0.37
Hearing-Imp. Children	4.0	0.0	0.00	5.0	0.0	0.00
Norm-Hearing Adults	3.8	0.4	0.20	3.4	1.3	0.60
Norm-Hearing Children	4.0	0.7	0.32	4.4	0.9	0.40
Average	4.0			4.3		
SE	0.05			0.33		

Table 5.10. Results for Experiment 2

5.4 Experiment 3

5.4.1 Methods and Procedures

The goal of this experiment was to study subject's motivation when watching visual stimuli using multimedia technology. For simplicity and to avoid fatigue in the subjects in this case there was no attempt to compare the effect of different stimuli, presented with the same methodology, on the subject's voice, but only how much they were motivated by such a methodology.

5.4.1.1 Selection of the stimuli

Two stimuli were chosen for this experiment. Both are digital video clips showing real images. The first one shows a jet aeroplane that is landing on a runway, seen from its side. The second one shows a tall building being demolished, collapsing in a cloud of dust (see Figure 5.18). Each video clip lasts about 20 seconds. These video clips were chosen for various reasons. Both are impressive, and in both of them one or more visual dimensions varies simultaneously to create the motion[1]. In the first video (the aeroplane landing) the feeling is that speed and height are decreasing. In the second video (the building) an heavy mass is falling down. Another reason for choosing these clips was that they were in a set of video clips previously shown to speech therapists who commented that those videos could be really good candidates for a particular type of application using visual feedback, since the "lowering" action is appropriate for correcting the tendency of many hearing-impaired speakers to increase their pitch and loudness.

[1] This is true for most of the length of the video clips. In the case of the aeroplane landing, the final impact with the runaway causes a quick change of the motion. In the case of the building collapsing the final cloud of dust suggest expansion. However these elements do not affect the aim of the experiment.

5.4.1.2 Implementation of the stimuli

The two sequences were taken from a library of video clips contained in a CD-ROM (Adobe Premiere). They were encoded using the Audio Video Interleaved (AVI) video codec (VIDS RT21) and played back using a MS Windows Media Player. The Media Player was set in "loop mode" so that at the end the video clip restarted from the beginning. Each video clip was repeated four times.

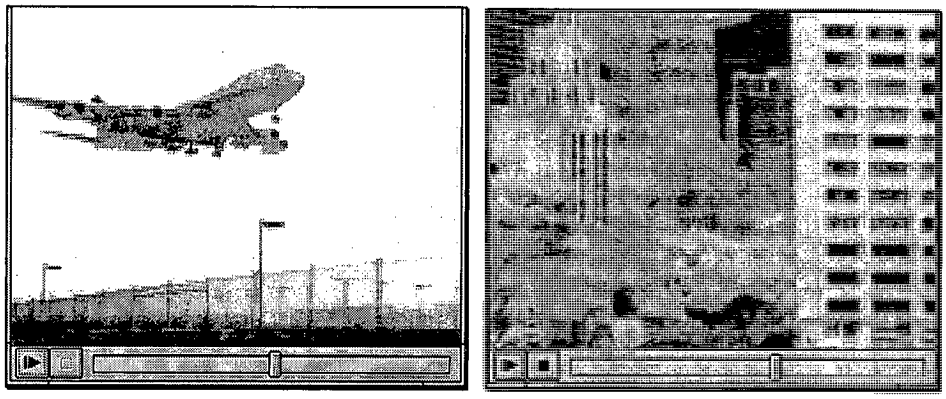


Figure 5.18. The two video clips used in Experiment 3

5.4.1.3 Subjects

Ten subjects attended the experiment, five of them hearing-impaired, and five normal-hearing, aged over twenty. All the subjects belonged to the group of people who participated in Experiment 1.

5.4.1.4 Structure of individual trials

The experiment was run after the Experiment 2. Each 20 second long video clip was shown four times. Again the subject was asked to try to accompany the images with their voice producing any sound or noise they considered appropriate, for the same reasons explained for Experiment 2.

5.4.1.5 Interviews

At the end of the presentation of the two video clips subjects were asked to rate on a form how much they liked them. As for Experiment 1 the form had five scores for each stimuli, "strongly disliked", "disliked", "neutral", "liked", "strongly liked".

5.4.2 Results

In the evaluation of results the following elements were taken into account: 1) the subject's score, from the form, and 2) the impression of motivation in the stimuli that the video recording of the experiment gave to the three independent assessors who helped in the evaluation of the results. This information was used in the actual implementation of the visual feedback for speech rehabilitation, described in Chapter 6. Figure 5.19 and Figure 5.20 show the results of the answers given by hearing-impaired and normal-hearing subjects to the question "how much did you like this video clip?". A score of 1 means "strongly disliked", 2 means "disliked", 3 means "neutral", 4 means "liked" and 5 means "strongly liked".

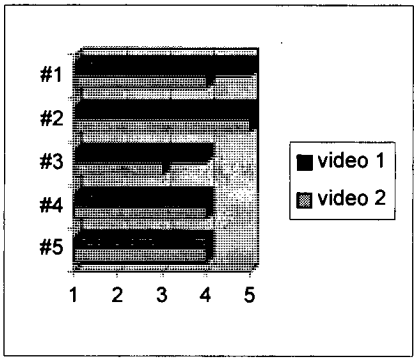


Figure 5.19. Scores of the multimedia stimuli for the hearing-impaired subjects.

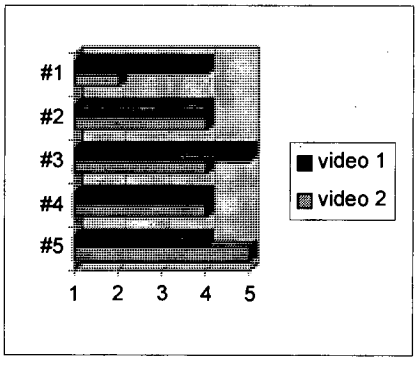


Figure 5.20. Scores of the multimedia stimuli for the normal-hearing subjects

5.4.3 Conclusion

As shown in Table 5.11 The visual stimuli using multimedia technology were positively rated by all subjects. Each subject gave a "liked" or "strongly liked" score for at least one of the two video clips, suggesting that this technology is worth investigating as a speech feedback methodology.

video 1			video 2		
Average	SD	SE	Average	SD	SE
4.4	0.5	0.24	4.0	0.7	0.31

Table 5.11. Average, Standard Deviation and Standard Error values of the scores for Experiment 3

However, this experiment was very preliminary in scope. In order to investigate the possibility of using this technology as a form of visual feedback for speech rehabilitation, further experiments comparing people's preferences for this type of technology against the more traditional methods of visual feedback, are needed.

CHAPTER 6

Design and Implementation of a Prototype System for Rehabilitation of Hearing-Impaired Speech

6.1 Introduction

In the previous Chapters of this thesis the need was highlighted for improved visual feedback for hearing-impaired speech rehabilitation and a method was proposed for achieving this improvement. The following diagram summarises the proposed method.

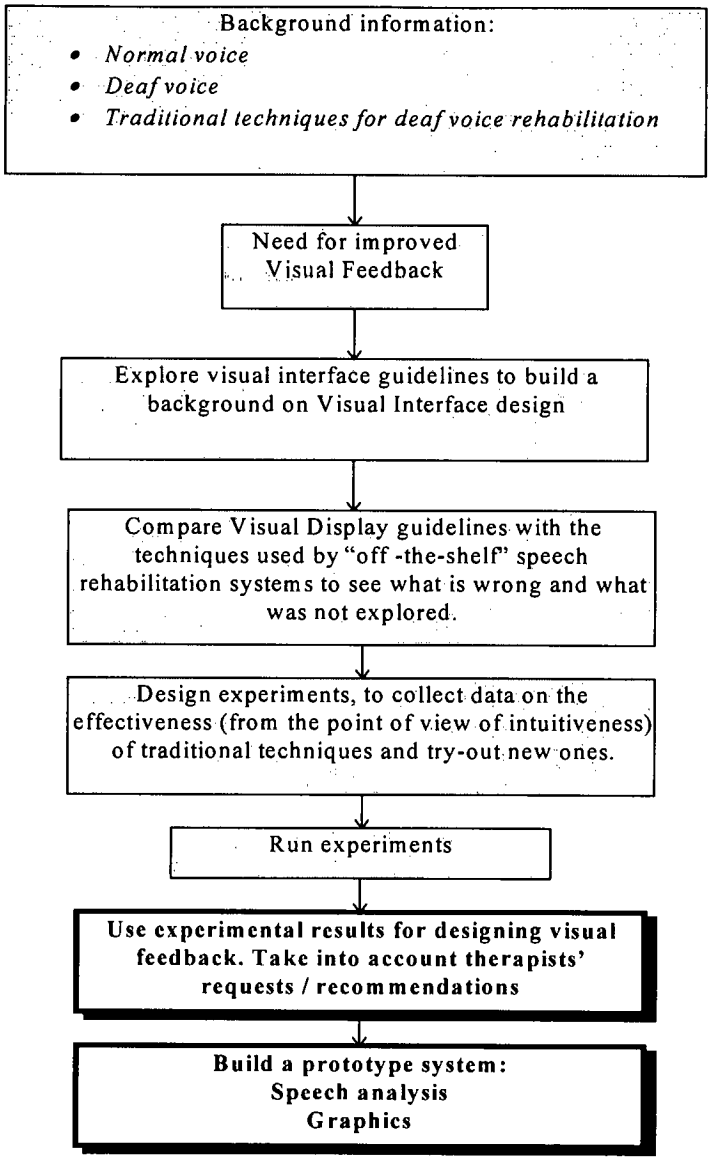


Figure 6.1. The proposed approach for designing visual feedback for voice rehabilitation of hearing-impaired speech

This chapter deals with the activities described in the last two blocks of the diagram:

- The design of appropriate visual feedback, using the results of experiments and integrating them with the therapists' requests and recommendations on effectiveness and motivation.

- The implementation of appropriate visual feedback in a prototype system, taking into account real-time speech analysis and display issues. This work covers two aspects: implementation of the speech analysis modules and implementation of the graphics modules.

6.2 Design of Appropriate Visual Feedback

In the design of visual feedback, three elements will be considered:

- the results of the experiments
- the therapists' recommendations
- the Visual Interface Design guidelines

To summarise, from the experimental data it has been shown that the visual stimuli using *dimension*, *distance*, *length* and to a lesser extent, *luminance*, are good candidates for designing visual feedback for *loudness*; some of the visual stimuli using *speed* and *line-graph* are good candidates for designing visual feedback for *pitch*; no conclusion can be made for specific vowels and consonants. Graphics showing a high level of realism were all well accepted.

Therapists' recommendations

The design phase of the prototype system took advantage of consultations with speech therapists, teachers of the deaf and also hearing-impaired users themselves. The results of the experiments were discussed with them in order to guide the development of the prototype system, and the development phase itself was carried out with a close interaction with them. Details of these consultations are reported in Chapter 7.

As discussed in Section 4.2, one of the major concerns expressed by therapists in relation to any form of visual feedback is negative reinforcement, as a consequence of both inaccurate analysis and because of the fact that visual feedback often tends to encourage users to shout or strain their voices in order to activate the most attractive part of the display. This is particularly true for loudness and pitch. Another concern is the frustration that a user may feel when the goal requested is too difficult to achieve, with consequent de-motivation. Furthermore, the display should be easily understandable, one of the major points of this thesis. Another recommendation from some therapists concerns the confusion that users sometimes experience when they are following a rehabilitation programme dealing with more than one speech feature. Displays dealing with the same type of speech feature (for example loudness) need to be characterised by some common style, that should be clearly distinct from the style of display dealing with some other speech feature (for example pitch). This separation of styles will help the user to "switch" from drills with one feature to another, reducing the danger of mixing-up the purpose of the different displays[1]. A simple colour / texture code may be adopted. For example, all feedback relating to loudness might use a blue background for the graphics screen, while all feedback related to pitch might have a wooden-textured background, and so on. The point in this

[1] One therapist said that she uses a different room in the speech laboratory for rehabilitating each different speech feature. The room-speech feature association makes it easier for the user to understand that, for example, the topic of the day is pitch instead of loudness.

case is not simply to find an appropriate style for each speech feature, but to differentiate clearly between them in order to make them more easily recognisable.

Recommendations of Visual Interface Design guidelines

Visual feedback should be designed to take into account rules of good visual design practice, such as the Gestalt theory (see Section 4.4.2), together with more specific guidelines regarding the functionality of the various sections of the display (for example Grether & Backer, 1972). In the case of visual feedback, the display should give space and importance to the feedback itself, avoiding filling the screen with elements that can distract the user. Accessory controls, such as the microphone sensitivity setting, the optional sound level meter (for monitoring and calibration), and other optional controls, should be put in a peripheral position to make it clear that these functions are used only for set-up purposes. If "start" and "stop" buttons are provided, they should be positioned close to the area destined for the feedback and they should be of adequate size, in order to make it clear that these buttons control the feedback itself. At the same the buttons should not create obstacles to the animation. If a "help" button is provided, it should be outside the main field, and also far from the auxiliary controls, since it provides a complete different function. Ideally all the displays should be consistent with these rules, and all the various elements should be kept in the same position throughout the various screens used, to quickly familiarise the user with them.

6.2.1 Loudness

Experiment 1 described earlier showed that visual stimuli using *dimension* are particularly suitable for stimulating changes in loudness proportional to the size of the object. In fact this association is already used in previously published rehabilitation systems. Visual stimuli using *distance* produced a significant change in loudness. Results concerning *luminance* exhibit a higher standard deviation, suggesting these stimuli may work with some subjects but not with others. All three of these modalities were considered for designing real-time visual feedback for loudness. Another modality which is appropriate for loudness was *length*, and this modality is used in all screens for loudness feedback, and as a display for the current voice level in the setting area of the screen (see Figure 6.2).

Therapists' recommendations

Therapists have complained that visual feedback of loudness (as well as pitch) sometimes encourages the user to shout, in order to activate the most attractive part of the graphics. Their recommendation is that visual feedback should attract the interest of the user mostly when the voice level is at a "medium" value, as measured by the therapist during the therapy session, and this value should be easily adjustable up or down in order to match the goal of the session.

Components of the visual display

In the system described here the display consists of the following elements:

- the visual feedback itself
- the "settings area", consisting of a microphone level setting, the switch to select a smooth or prompt response, and a display of the current sound level
- the "help" button
- the "exit" button

The visual feedback itself, being the only element of interest for the user, occupies the most of the screen, taking at least 70% of the available space. The settings are meant to be used by the therapist only, and are small and not distracting for the user, but at the same time easily accessible while the user is speaking without the need for transfer to another screen. An appropriate position for these controls is an area near the bottom-right of the screen. The "help" button is rarely used, and although it may be exercised during the first few times that the system is used, it will subsequently be used only occasionally after that. In consequence it is best to always put the button in the same position, for familiarisation, and not too close the other controls. The top left corner is reserved for this function. In most displays the help function is implemented in two ways: a traditional method where pressing the help button activates a "help page" containing text and possibly diagrams; and a "video help" facility, where explanations are given by means of a video clip showing a close-up of a person, with optional captions, or a person using "sign language". The help button is actually divided into two sections, labelled "help" and "video help", whose functions are clear (the "video help" function will be discussed in Section 6.2.5). The "exit" button is only used at the end of the session, and is positioned at the bottom-right, which is conventionally considered the "end" of the screen.

Using *Dimension* as a feedback for loudness

Loudness is a speech feature more likely to be practised by young users, at an early stage of the rehabilitation process. The following is an example of a display attractive to children: a friendly character (a dog, see Figure 6.2) is constantly moving its tail, looking at the user, and waiting for the user to speak.

Figure 6.2. Using *Dimension* as a feedback for loudness

The speech level controls the size of the bone, very small in case of silence, and bigger and bigger while the speech level increases. As the bone becomes bigger, the dog looks more and more interested in it, by enlarging its eyes and by sticking its tongue out of its mouth. When the bone reaches the "target" dimension, the dog looks very happy and "barks" (through a caption appearing on the screen). As the level of the user's voice increases, the bone becomes even bigger, but at the same time turns more and more on its side, away from the dog, which looks less and less interested since the bone is no longer within reach. Note that the setting area on the bottom-right of the screen uses an horizontal bar to monitor the voice level. This choice was based on the results of Experiment 1, where the *length* stimuli had a relevant link with loudness. In this way, the length of the bar is a reinforcement of the main feedback obtained using a change of dimension.

For details about the tools and programming languages used to implement this display, and the displays presented in the following pages, see Section 6.3.3.

Using *Distance* as a feedback for loudness

The display of Figure 6.3 was designed in order to be very easy to understand: when the user is silent a small character's face is displayed on the screen. The character has a hand near his ear, suggestive of someone who cannot hear because the voice level is too low (see Figure 6.3). As the user speaks, the face becomes bigger, and the hand near the ear is gradually removed. When the voice level reaches the "target" level, the face looks happy, showing that the character can clearly hear the voice. If the voice level increases further, the character's face becomes bigger and shows annoyance to the excessive noise, covering both ears. To make things even clearer, the character "speaks" through captions, giving a textual feedback like "I can't hear you", "Now it's fine", "It's a bit too aloud".

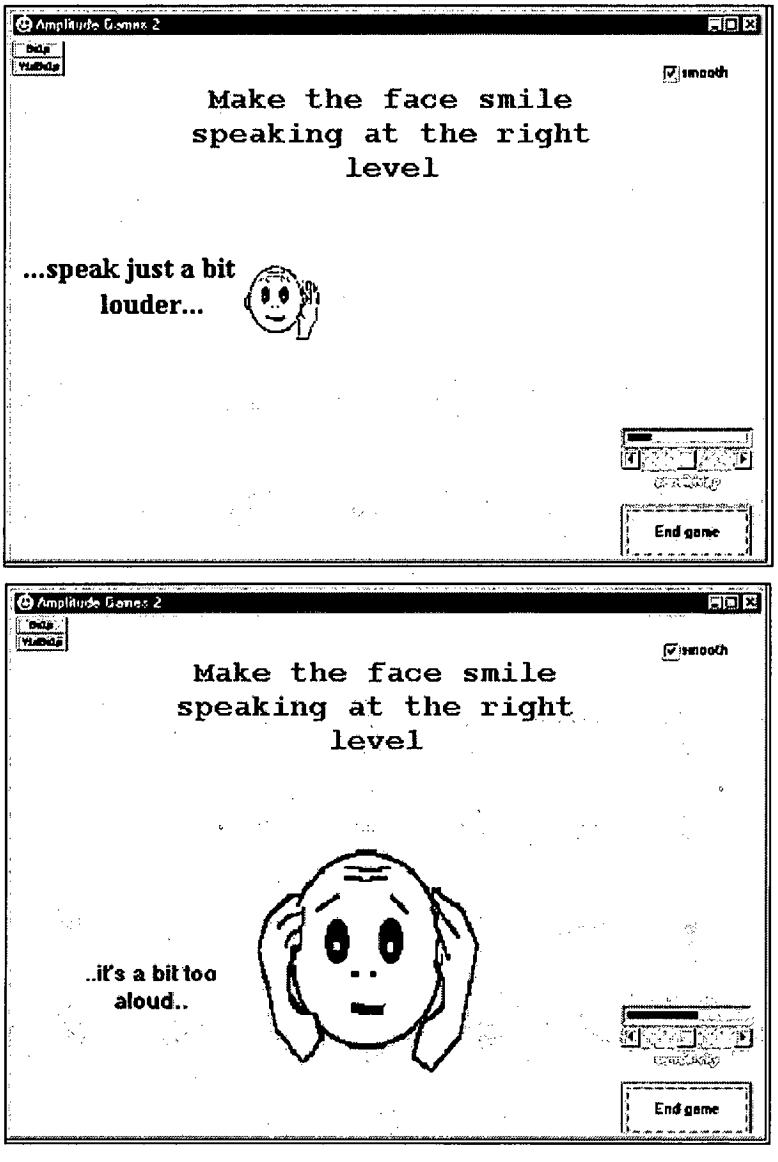


Figure 6.3. Using *Distance* as a feedback for loudness

Using *Luminance* as a feedback for loudness

Visual feedback using a change of *Luminance* was also designed. A ghost appears out of the dark as the voice level increases, becoming more visible and less blurred. When the voice reaches the "target" level, the ghost is clear and its eyes become red. If the voice level increases further, the images become even brighter but also blurred, until the ghost is no longer visible (see Figure 6.4).

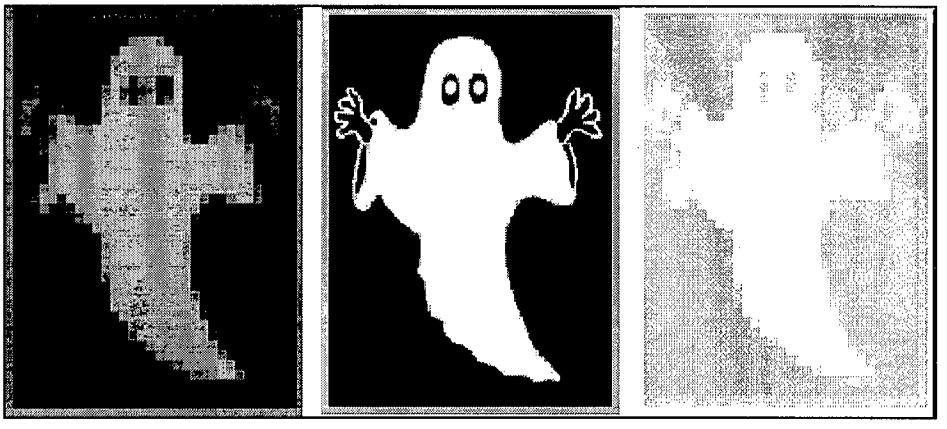


Figure 6.4. Using *Luminance* as a feedback for loudness

6.2.2 Fundamental frequency

Results from the experiments indicates a connection between the speed of a smooth spinning object and the change that this stimuli produces in the fundamental frequency of a speaker asked to accompany such images with the sound that in their opinion is the most appropriate. Therefore a visual feedback using this modality was designed, since it appears to be a promising method for showing pitch. Nevertheless, the *linegraph* already used also produced good results indicating that it is an effective visual feedback for displaying fundamental frequency. Visual feedback using the salient feature used in the *realistic* stimuli was designed too, since from Experiment 1 this modality appeared worthy of investigation. In the following discussion the terms "fundamental frequency" and "pitch" are used synonymously, according to widespread (but strictly inaccurate) terminology.

Therapists' recommendations

The introduction of new types of visual feedback for pitch rehabilitation is viewed both with interest and concern by therapists: interest because a new method can potentially be more effective than previous ones: concern because the new method may change considerably the way intonation is taught. The traditional association 'high pitch = high position, low pitch = low position' may take some time to be understood by users, but once this is achieved it is an effective method for representing pitch versus time, in the same way that language teachers explain stress and intonation to

normal-hearing people using lines drawn over words and phrases. Representing pitch with a linegraph is therefore the link to a more advanced stage of speech rehabilitation. In order to match these two contrasting aspects, a novel visual feedback, based on more intuitive empirical evidence than the traditional one using linegraphs, and hence probably more effective in an early stage of speech rehabilitation, is designed to be faded in gradually, giving more weight to the traditional one as the rehabilitation process advances.

Recommendations of Visual Interface Design guidelines

Visual feedback for pitch generally requires more elements on the screen than feedback for loudness, since often *time* needs to be displayed as well as pitch, possibly needing more controls and options for handling this additional dimension. More careful "grouping" of elements is then required. In certain cases the use of "pull-down menus" is appropriate for grouping commands or settings that can be listed under one or more categories of similar functions. However pull-down menus hide possible options, and are not always appropriate. When these options need to be visible on screen at all times for more immediate selection, icons representing the function they control, are a better choice. When commands like "start", "pause" and "stop" are required, it is appropriate to group them together from left to right, following the conventionally accepted temporal sequence. The "exit" button should follow the sequence, but be separate from the others, since it is used only once at the very end of the session. A bottom right screen position is recommended.

Using *Speed* as a feedback for fundamental frequency

The experiments described in Chapter 5 showed that the display of a round object, with a smooth surface, spinning at different speeds represents an intuitive method for giving feedback on fundamental frequency. In order to make this feedback more clear, the object can be textured in a way that emphasises the perception of speed, avoiding "tiled" or chequered patterns which may create a confusing stroboscopic effect and give the appearance that the object is steady, or spinning in the opposite direction at certain revolution rates[1].

[1] A texture able to visually accommodate a high dynamic of revolution rates is the spiral. It is often painted at the centre of aeroplane propellers because of this property.

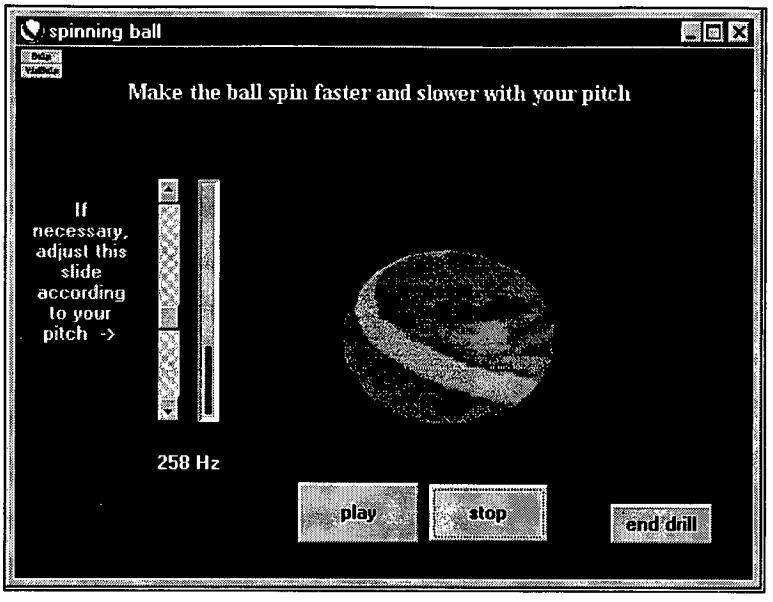


Figure 6.5. Using *Speed* as a feedback for fundamental frequency

The feedback may be used "as is" in a preliminary stage (see Figure 6.5), and after that it may be used in a multi-modal display together with the traditionally used line-graph technique (in this case the spinning object moves also to the right with time, and up and down with pitch, following the line-graph edge). Multi-modal displays are generally not advisable as a feedback for speech rehabilitation because they are too complex to be processed by the user (as discussed in Chapter 3); however, in this case the two dimensions (speed and position) change in concert, giving redundant information on the same speech feature. This should achieve the goal of a smooth transition between one method and the other, and may have also the effect of speeding-up the process of understanding the less intuitive one. However this type of display was not considered, since it needed to be based on the results of the display shown in Figure 6.5 and its implementation is left for future enhancements in this field.

Using *a linegraph* as a feedback for fundamental frequency

A pointer, moving constantly to the right with time, and up and down with fundamental frequency, and leaving a track (or path) over the screen is the most widely used method for displaying pitch. Therefore this method was used as a basis for various visual feedbacks for pitch, trying to enhance those already seen on other systems by considering therapists' suggestions and visual design guidelines. Two kinds of display were designed, one for real-time feedback and the other for delayed feedback of fundamental frequency. The real-time feedback is designed to be used with continuous sounds, such as long vowels, and is more likely to be used by people in an early stage of speech rehabilitation. The delayed feedback display is designed to be used with words and phrases, and should be able to show fine detail of fundamental frequency changes.

Real-time feedback

Real-time pitch feedback, due to its highly interactive nature and the presence of running time as one of the dimensions being displayed, is more suitable than delayed feedback display for inclusion in a video-game setting. The advantages and the related problems of this kind of approach were discussed in Sections 3.2.1 and 4.2. Essentially, it is known that a video-game approach is highly attractive especially for young users, but therapists are concerned about the danger of frustration in users who cannot easily reach the goal of the game. A solution for this problem is to design visual feedback that, apart from its main goal of being effective in speech rehabilitation, is also fun to use, whether or not the user is able to achieve the goal. A tentative balancing of these issues was attempted in the design of the "pitch aeroplane" (see Figure 6.6).

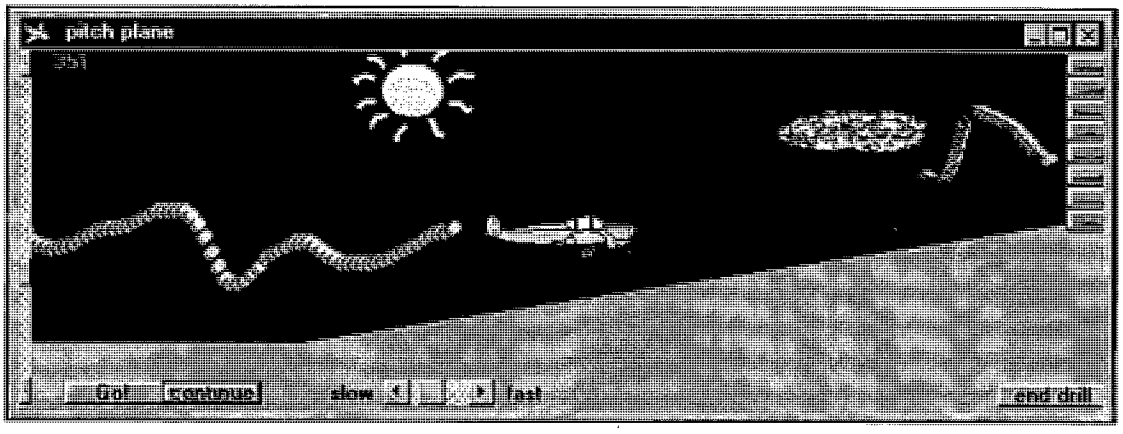


Figure 6.6. Using *a linegraph* as a feedback for fundamental frequency

In this feedback, user voice's pitch controls the height of an aeroplane, which moves constantly from left to right. The aeroplane flies on a background which can be easily changed by the therapist or the user themselves. Some of the backgrounds are very easy to follow, showing a straight panorama with no changes in height of the objects shown. Others add some obstacle that the user has to avoid, or in some cases, to hit, depending on the situation that the therapist wants to create. The aeroplane never actually stops, even if it crashes into obstacles, so as to allow the user to complete the screen without being penalised for errors. If the plane hits the ground or other objects, it flashes with an animation until it is free again. More complex backgrounds allow practice of different intonation patterns. The therapist can also draw a specific background using the mouse, in order to create new obstacles and intonation patterns (as in the right hand side of Figure 6.6).

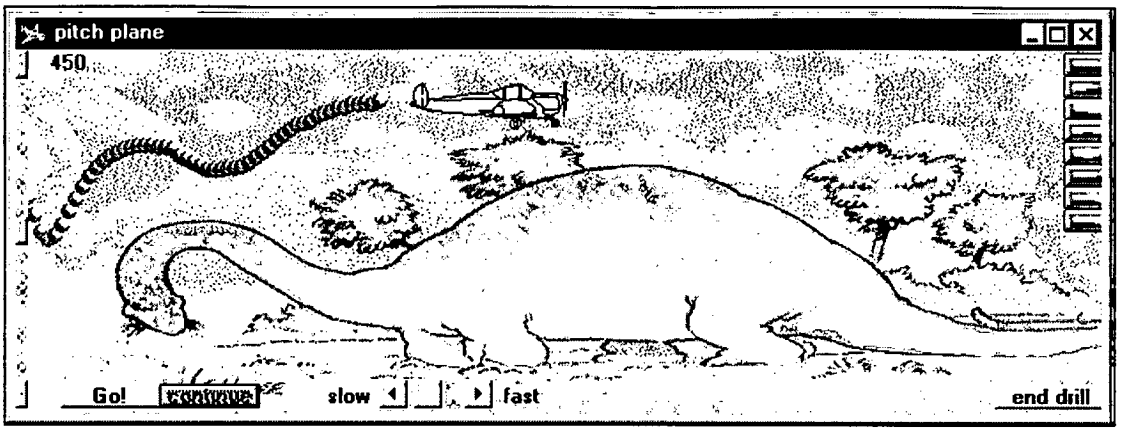


Figure 6.7. Another display using a *linegraph* as a feedback for fundamental frequency

Alternative intonation patterns can be practised with a variety of backgrounds with different pictures (see Figure 6.7), useful for helping the therapist to stimulate user interest.

Delayed feedback

Delayed feedback is useful when detailed analysis of pitch is required, for example for refining intonation in words or phrases. The system here waits for an utterance, and display the pitch track as soon as possible after the end of the utterance. The display allows the possibility for comparison between the user's performance and an utterance produced by the therapist or other user, or taken from a database stored on the computer's disk. For this reason an easy and quick way to "keep" and "re-display" significant utterances is provided.

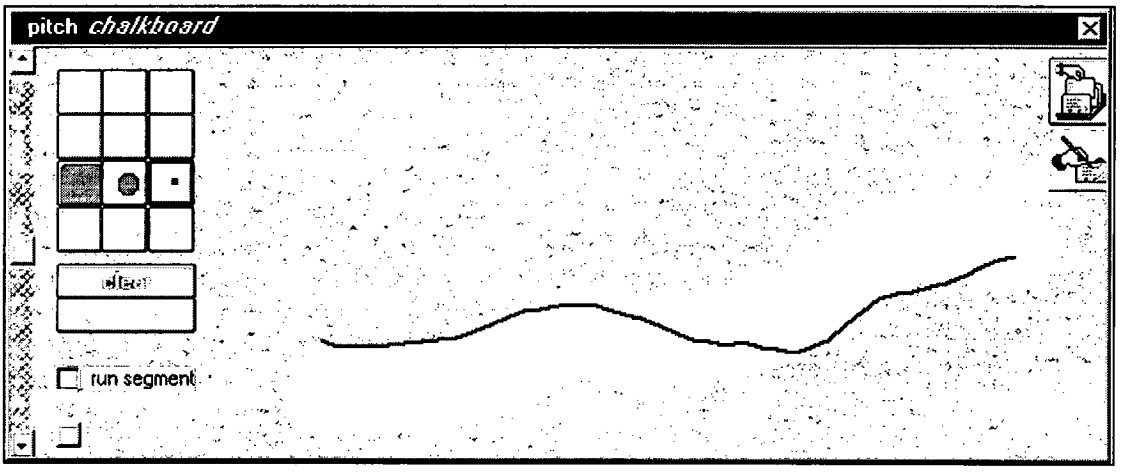


Figure 6.8. Delayed feedback for fundamental frequency

In order to give the feedback the feeling of a familiar object, it was designed to resemble a blackboard. A selection of chalks, with different colours and sizes was realised as a set of small icons

grouped on the left side of the screen. The left side was chosen, following the temporal sequence convention, since the analysis is activated by the action of selecting a chalk (a "speak now" indication appears as well at this time). Different colours help to present different tracks at the same time. An "erase" button clears the board, while a "keep the last" button stores the last track for persistent display. The different sizes of the chalks allows the therapist to draw (by using their voice) a reference track, wide enough to allow reasonable margins for the user to overlap with their attempts, using a thinner chalk (see Figure 6.8). The display uses only the upper half of the total screen area, in order to leave the lower half for a second instance of the feedback to be open at the same time. This gives the possibility of displaying a reference pattern on a separate screen rather than using the same one, avoiding too strict a comparison of the reference and user's tracks.

Using *Perspective* as a feedback for pitch

Experiment 1 showed a significant link between the speech feature of *pitch* and some of the stimuli belonging to the set named *realistic stimuli*. Furthermore, Experiment 2 and 3 showed that images combining 3D and motion are highly motivating. A visual feedback was designed according to this finding. This visual feedback associates pitch and height in a realistic simulation of the flight of a bird (or helicopter), where the distance from the ground is controlled by the pitch of the speaker. Figure 6.9 shows some frames from the display.

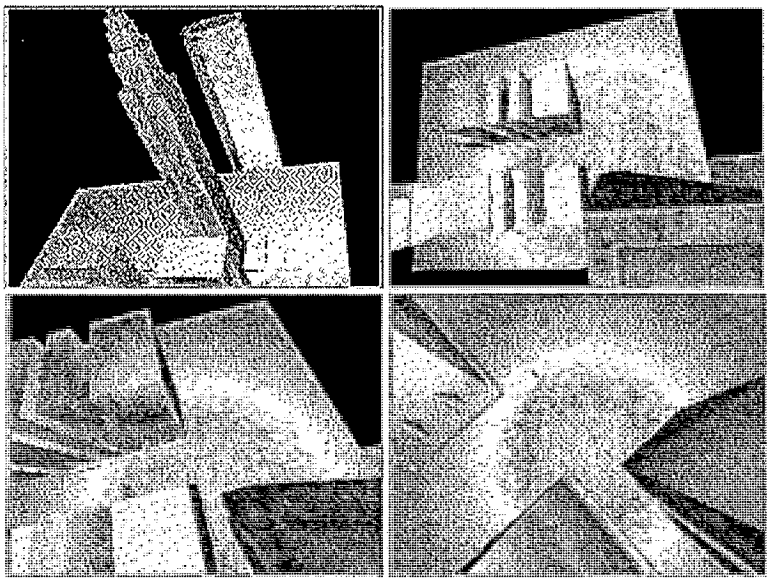


Figure 6.9. Using *Perspective* as a feedback for fundamental frequency

6.2.3 Vowels

The experiments gave no significant results for an intuitive link between vowels and the visual stimuli chosen. Therefore the design of the visual feedback will consider only the analysis of previously described systems, therapists' comments on them, their recommendations, and some guidelines on Visual Display Design.

Therapists' recommendations

Vowel feedback often uses categorical displays, which give no indication of how far from correct settings a vowel production is. Moreover it is important to find a method which can help users to know how to improve an incorrect vowel. Also, the goal of the session should be easy to achieve, to avoid frustration. It is likely that many hearing-impaired users will never be able to produce the complete set of vowels in the same way that a normal-hearing speaker does. Therefore it is better to avoid confronting them with targets which are too difficult to reach.

The symbols used for representing vowels should ideally follow standard conventions. For example the International Phonetic Alphabet (IPA, International Phonetic Association, 1993) or the Machine Readable Phonetic Alphabet (MRPA, Harrington et al., 1996) should be incorporated. However therapists may use different methods with different users (for example IPA, MRPA, or a mixture of the two). A set of symbols should then be provided to cope with the different cases.

Recommendations of Visual Interface Design guidelines

The feedback needs various settings for user sex and language. Recommendations involve use of functions which group menu items that do not need a "pictorial" description, but that can be "hidden" under labels that categorise them clearly. In this case: sex (child, female, male), language / accent (RP English, Scottish, French...), etc. Furthermore, feedback gives therapists and users the opportunity to adapt targets according to specific user needs. This involves a more advanced familiarity with Graphic User Interface conventions, especially regarding the use of the mouse. These conventions may vary with the computer system used. For example, the same "drag-and-drop" function may be implemented differently under the MS-Windows environment, the Mac-OS environment or X-Windows. It is advisable to design the feedback so that these interactions are achieved by means of procedures which are consistent with the environment in use, assuming that the person who is operating the computer system is already familiar with the conventions of that particular system.

Using a pointer in a normalised plane as a feedback for vowels

The idea on which visual feedback for vowels is based is the representation of the quality of the vowel produced by the user in real-time on a two-dimensional "vowel space". A pointer moves in a continuous (versus categorical) mode indicating the vowel quality. The vowel space is normalised so as to approximately represent the position of the tip of the tongue in the mouth (the Sydral and Gopal normalisation method is used, as described later in this chapter.) Therefore, to correct a wrong vowel the speaker can try to move the tongue onto a target by moving the tongue in the direction indicated by the display. Three different displays are available:

1. the "reference" display, showing the reference vowel position for the chosen language / sex combination (see Figure 6.10);
2. the "movable targets" display, where vowel position is movable (using a "drag-and-drop" action). The position is chosen by the therapist to make the goal easy to reach (often hearing-impaired people have a restricted vowel space, where all the vowels tend to cluster together around the mid vowel /ə/), or for focusing only on the vowels of interest by moving away the others (Figure 6.11);
3. the "vocal tract" display, showing the vowel produced by the user as a pointer inside a cross section of the vocal tract. No reference vowel is shown here, however the therapist has the option to drag and drop labels to mark positions (see Figure 6.12). The normalisation method used in this display is the same as the other two, but the picture of the vocal tract in the background helps the user to understand how to correct tongue position in order to reach the desired target.

Diphthongs are dealt with in the same way as steady vowels, and the result is a movement of the pointer from the vowel's initial position to the vowel's end position. A better solution may involve the use of a delayed feedback display, since it is not common practice to teach diphthongs in isolation.

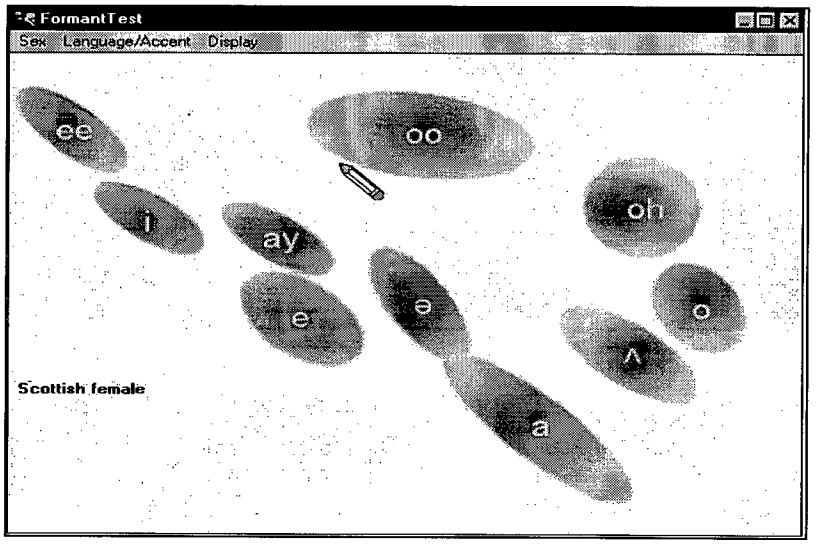


Figure 6.10. Vowels *reference* display

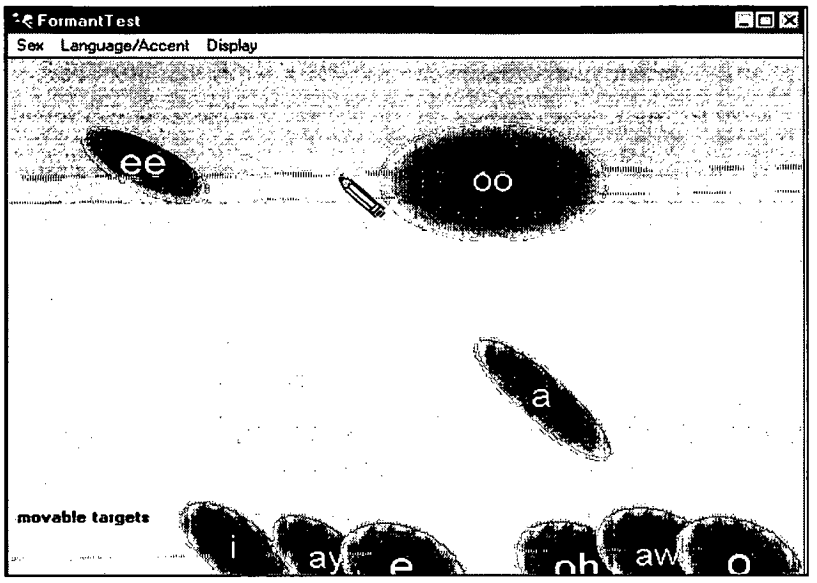


Figure 6.11. *Moveable targets* vowel display

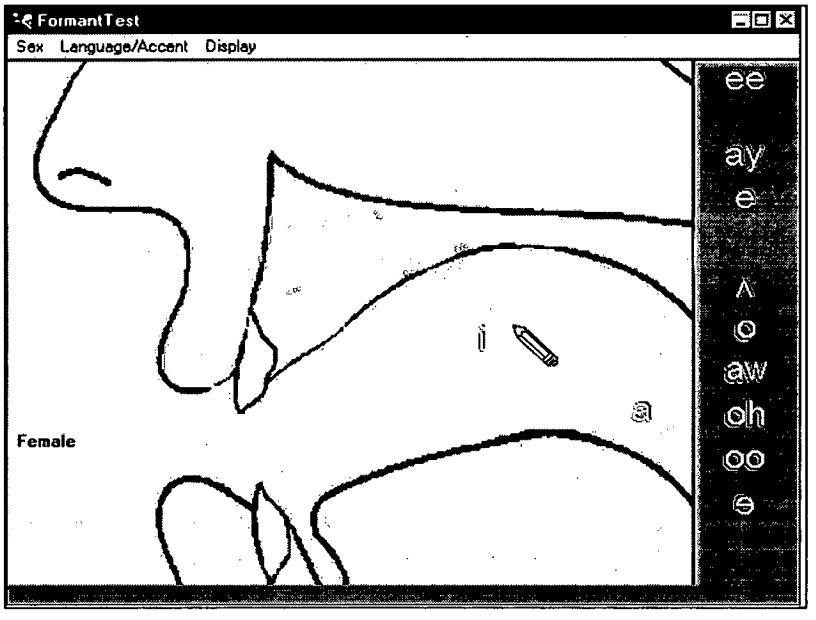


Figure 6.12. *Vocal tract* vowel display

6.2.4 Consonants

As for results from vowels, experiments gave no statistical evidence for an intuitive link between consonants and the visual stimuli selected. Several feedback systems for consonants have been considered, focusing on some of the problems with consonants that have the highest priority in speech rehabilitation. The designs were based on the analysis of published speech rehabilitation systems, comments from therapists, and visual display guidelines. The first visual feedback deals with the contrast /s/-/ʃ/ (like in "sea" - "ship") that can be handled efficiently with a real-time display, allowing the user to practice with long consonants, and see immediately on the display the effect of changes in articulator positions.

The second feedback deals with the contrasts /t/ - /s/, or /b/ - /s/ (for example in words like "tea" vs. "sea", "bell" vs. "spell"). Plosive sounds are best practised at the beginning of words, or embedded into words, so a delayed feedback using a speech segmenter will be optimal in this case.

Therapists' recommendations

Therapists welcomed with enthusiasm the idea of showing an animation of the movement of the tongue while producing voiceless fricatives such as /s/ and /ʃ/, since the narrow opening of the lips hides the view of the tongue position. Some published systems use video disks shown on a TV screen for teaching tongue movements associated with different consonants. But if the animation could be controlled by the actual speech production, this would be a powerful feedback mode. Furthermore, as for displays for vowels, the feedback should give a continuous representation of the sound, rather than be categorical, ideally with indications on how to correct a wrong production.

Recommendations of Visual Interface Design guidelines

The /s/ - /ʃ/ consonant display may be realised using a photographic quality image of the cross section of a real vocal tract (X-ray image), or a schematised drawing of the same. As discussed in section 4.4.4., the second solution is preferred, since the real image, full of unimportant details, is likely to be less clear and immediate to understand than a well selected drawing, which is able to highlight the feature of interest and take advantage of appropriate use of colours and display of details only where they are needed.

Using a vocal tract animation as a feedback for the /s/ - /ʃ/ contrast.

As shown in Figure 6.13, the major feature of the display is a sagittal cross-section of the vocal tract, with the profile of a face. The tongue in the animation moves following the user's consonant production, and the letters "s" or "sh" are depicted accordingly. The air stream is shown as white dots moving quickly through the vocal tract and outside the mouth. The right side of the display contains some auxiliary information for the therapist: two "needle-type" indicators, one showing the "centre of gravity" of the speech production spectra (as explained later in the section describing implementation), the other showing the speech level. The numeric value of the centre of gravity is also displayed.

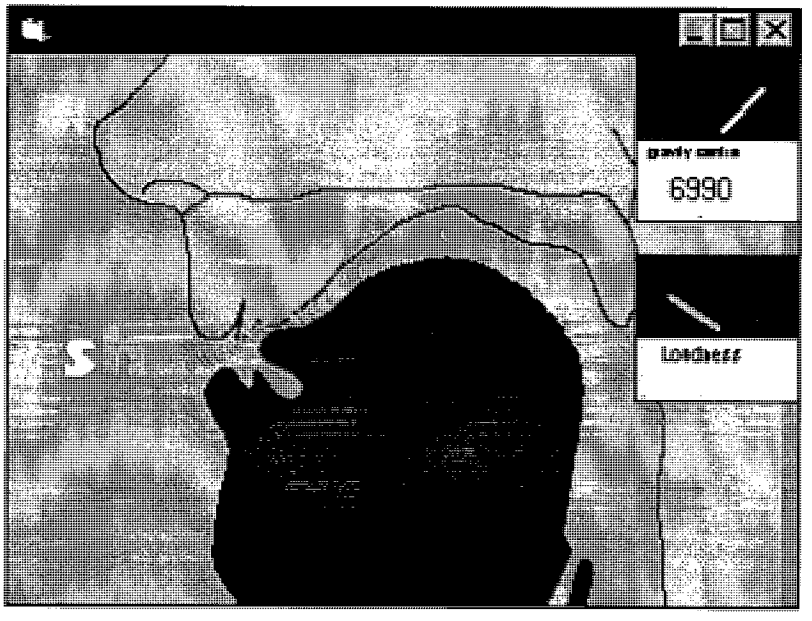


Figure 6.13. Using a vocal tract animation as a feedback for the /s/ - /ʃ/ contrast

Using a delayed feedback for consonant contrasts embedded in words

An automatic phonetic segmenter, is used to label the different sounds in a phrase or word. When this capability is used for analysing contrasts in initial and final consonants, it is difficult to design visual feedback to give a non-categorical result, unless the model used is so accurate (and therefore complex) as to be able to categorise the many variations of the sounds of interest. Since this possibility was not achievable with the actual segmenter being used, a range of categorical displays were designed. These were eventually not used, since adequately robust phonetic models, able to accommodate many variations in order to give a less categorical feedback, could not be built.

Using animation for teaching consonant contrasts

Animation is an excellent method for displaying dynamically the position of the tongue and the jaws in consonant production. Since the perception of consonants relies on hearing high frequencies, and most hearing impairments are accompanied by a loss of sensivity in this frequency range, consonants appear completely silent to hearing-impaired people with even mild deafness. A set of animations was designed, covering the most common problems in consonant production: the /k/ sound (as in "key"), the /s/ sound (as in "sea"), the /t/ sound (as in "tea"), and the contrast between /s/ (as in "sea") and /ʃ/ (as in "shoe"). Each animation can run continuously, showing the movement of the articulators again and again, or can be paused and controlled through a "slide bar", allowing detailed study of selected sections of the movement (see Figure 6.14). Some of the animations have an additional window with a real speaker shown from the front, and speaking in sync with the animation. In this way all the relevant information from the outside and the inside of the mouth are shown, making it easier for hearing-impaired users to try and imitate the speech production.

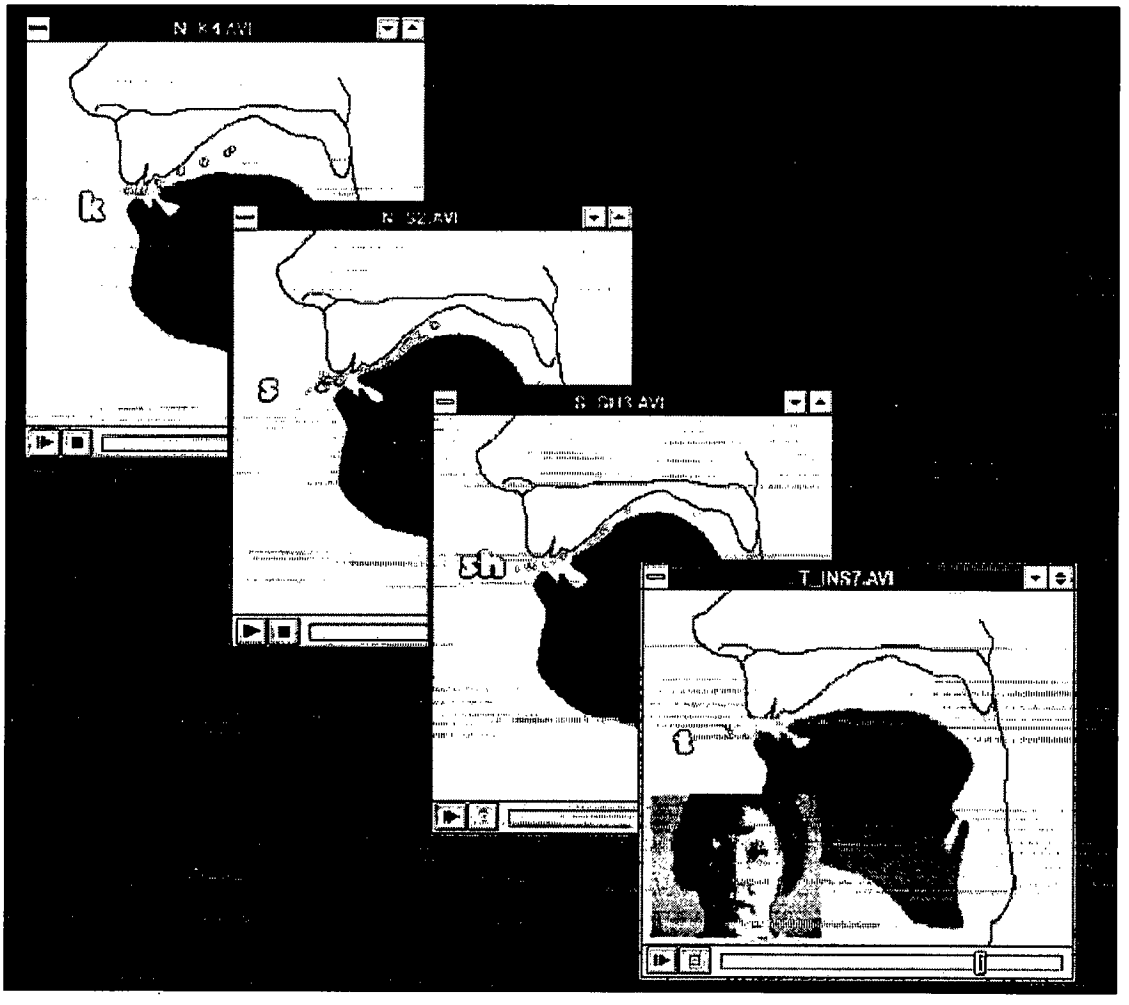


Figure 6.14. Using animation for teaching consonants and consonant contrasts

6.2.5 The Help system

A help system was designed, in order to make it clear for therapists and users how each feedback option works. As explained in Section 6.2 and 6.2.1, in most visual feedbacks a "help" button is provided, giving context-sensitive support related to the visual feedback being used. The help is provided in two ways: normal help and video help.

The normal help uses text, images and hyperlinks, (underlined words used for jumping to another position or page of the help system). The goal of the help system is to explain how to use visual feedback and to give supplementary information, (such as useful lists of words to pronounce).

The video help (see Figure 6.15) uses images of a real speaker, with optional captions, or a sign language interpreter. The user can choose either method according to their preference. Both versions have an audio track with the voice of the speaker that may be useful for therapists and for users with some residual hearing.

Figure 6.15. The two versions of the video help: lip reading (left) and sign language (right)

6.3 Implementation of Appropriate Visual Feedback

6.3.1 Host Platform

The possible alternatives for hosting the prototype system were Intel based systems, and Apple Mac systems, 680X0 or Power PC based systems. Other potential hosts, such as DEC Alpha, Silicon Graphics, Sun and other UNIX platforms were immediately ruled-out on the basis of cost and complexity for the therapist, clinic or user.

Although the Apple Mac range offers a more advanced platform for multimedia applications, it remains a less open architecture than the Intel based machines. Furthermore, in the view of usability evaluation carried-out in speech rehabilitation laboratories, initial surveys showed that both therapists and patients are more likely to have experience on a IBM-PC compatible machine.

The availability of the Intel Pentium™ processor at the start of the implementation stage, at increasingly attractive prices, and the accompanying availability of extensive multimedia software, combined with Intel's familiarity, dictated that it be chosen as the host platform for the prototype system.

6.3.2 Speech analysis

The task of visual feedback for hearing-impaired speech rehabilitation is to give the user a visual substitute of speech. The initial stage of this task is performed by speech analysis algorithms, which convert the speech signal into some numeric representation of the feature of interest. This numeric representation is then used to control some appropriate graphics.

6.3.2.1 Implementing Speech analysis on a PC platform

In order to speed-up the process of developing and testing the various speech processing algorithms, they were implemented initially on a plug-in digital signal processing (DSP) accelerator card (see Appendix B for an example of source code for the DSP), and then ported onto the main processor of the personal computer. This development process was faster because the board allows development of processor intensive routines to be carried out rapidly, without the requirement to optimise the code for speed at an early stage, and the development environment was independent from restrictions due to the PC's operating system. As an example of this, the numerical accelerator card was used to implement the Shafer-Vincent pitch tracking algorithm (Vieira, 1996) and to allow its testing in real-time, prior to its porting to the Pentium processor.

The numerical accelerator used was a Loughborough Sound Images LSI DSP32C card, with the following characteristics:

- AT&T DSP32C floating point processor
- Hardware Floating Point Operations
- 80ns cycle type
- 512K SRAM (70 ns)
- 32K SRAM (25 ns)
- Double 16 bit ADC/DAC Burr-Brown PCM converter, maximum sample rate 160 kHz
- 4th order Butterworth analogue anti-aliasing filters
- AT&T 'C' compiler and assembler/linker
- Shared memory with the host PC.

The software was eventually ported onto the main processor of the PC, a Pentium running at 90 MHz. The Pentium processor delivers superscalar[1] second-generation RISC performance. Its floating-point unit, much improved in comparison with the performance of the previous 486, and its improved fixed-point multiply instruction, together with a sophisticated cache featuring branch prediction, allow many speech processing algorithms such as filtering and fast Fourier transform (FFT) to work in real time[2], thereby avoiding the need for an expensive numerical accelerator card.

Audio input

Audio input is achieved via a device-independent interface to computer-audio hardware, based on the application programming interface (API) of Microsoft Multimedia Extensions 1.0. This allows the system to work on any Windows compatible audio system[3]. The application communicates with the audio hardware using the "Low-Level Waveform Audio Services" family of function calls. By means of these functions the audio card may be programmed with various sampling rate frequencies, and

[1] A superscalar processor is one that can fetch, execute and complete more than one instruction in parallel. By implication, a superscalar processor has more than one pipeline. In the processor pipeline, the execution of each instruction is divided into a sequence of simpler suboperations. Each suboperation is performed by a separate hardware section called a stage, and each stage passes its result to a succeeding stage. Normally, each instruction only remains in each stage for a single cycle, and each stage begins executing a new instruction as previous instructions are being completed in later stages. Thus, a new instruction can often begin during every cycle. Pipelines greatly improve the rate at which instructions can be executed, as long as there are no dependencies. The efficient use of a pipeline requires that several instructions be executed in parallel, however the result of any instruction is not available for several cycles after that instruction has entered the pipeline. Thus, new instructions must not depend on the results of instructions which are still in the pipeline. (source: Mips Technologies, Inc).

[2] Many instructions have sped up from the 486: JMP (from 3 to 1 clocks), CALL (from 3 to 1 clock, direct addr), RET (from 5 to 2 clocks, direct addr), PUSH (from 4 to 1 clock, mem) and MUL (from 13/26 to 11 clocks, unsigned).

[3] Such as SoundBlaster, Turtle Beach, or many others.

selected for 8 or 16 bits/sample. The sampling frequencies of 10 kHz and 20 kHz were used in the speech analysis. The lower sampling rate (10 kHz) is adequate for pitch tracking and formant tracking. The higher rate (20 kHz) is used when consonant analysis is required. In both cases 16 bit samples are used. A common data acquisition mechanism was designed for use with all speech analysis modules, with the goal of avoiding loss of data even in case of CPU overload. A double buffered, high priority routine assured that the speech samples were copied into a large first-in first-out (FIFO) buffer, capable of storing about 8 seconds of speech. The speech analysis modules and the graphic routines were running at a lower priority, sharing the remaining CPU time. Some of the speech analysis modules were used for delayed visual feedback, while others for controlling real-time visual feedback when continuous speech (such as long vowels) was produced by the speaker. This particular type of speech does not require an analysis of each single frame of the signal, because of the very slow change in speech features. For these reasons, if the time required for processing a speech frame and for updating the display of the visual feedback is longer than the speech frame itself, one or more frames are allowed to be dropped from the analysis, in order to keep the visual feedback in synchronisation with the voice. With these conditions, a maximum delay of about 150 ms. between a speech event and the resulting change in the feedback is achieved, with no adverse consequences.

6.3.2.2 Problems in analysing hearing-impaired speech

In Chapter 3 it was shown how the speech of deaf people is different from normal speech. These differences may create problems when deaf speech is analysed using the same analysis techniques that are used for normal voice. The consequence is an inaccurate presentation of the feedback.

As was previously noted, it was decided to use the least invasive technique possible as an input to the voice analysis process, that is a single microphone. This makes acoustically based systems a more practical option for many users, both in the clinic and at home. Many users are already familiar with microphones, and therefore less intimidated by the prospect of interacting with the system. For these reasons, therefore, acoustic signals have in general been the main choice of published systems (Bernstein, 1989). The main drawback of using solely acoustic input is that a number of speech parameters such as nasalisation, consonant production and laryngeal quality are not easily obtainable from the acoustic speech waveform. Systems that provide feedback on these parameters must use complex and sophisticated analysis techniques, (usually unreliable) or supplement the acoustic information with some physiological data. A second disadvantage of acoustically based processing, which even sophisticated analysis may not overcome, is that different articulator gestures may result in the same acoustic signal, making it difficult to infer articulator patterns with complete accuracy from the acoustic signal alone. Physiologically based analysis may overcome some of these problems by providing parameters which relate directly to the processes of speech production themselves, using

special sensors. Examples of physiological measurements employed in speech training aids include airflow, measured using a pneumotachograph (e.g. the CISTA / Panasonic training aid, Yamada & Murata, 1991, and the Gallaudet system, Ferguson, Bernstein, & Goldstein, 1988); nasal vibration, measured using a contact microphone or accelerometer attached to the nose (e.g. CISTA and the Visual Speech Apparatus, Arends, 1993); laryngeal activity, measured with a throat microphone (e.g. CISTA), electroglottograph (the Gallaudet system) or electrolaryngograph (Visual Speech Apparatus); and tongue contact patterns, measured using dynamic palatology (e.g. the CISTA / Panasonic aid). Perhaps the most sophisticated physiological aid is the Optopalatograph (Fletcher et al., 1991, Wrench et al., 1996), which monitors tongue-palate contact, tongue position within the mouth and lip and jaw movements, using a series of light-detecting sensors and dynamic palatography.

The disadvantages of physiologically based systems are however considerable. The equipment required to obtain the measurements is often delicate, expensive and difficult to adjust. Home use of this equipment is impractical. In addition, some of the techniques are relatively invasive when compared with acoustic measurement, and there is resistance to their use from both users and therapists. A final problem is that the measurement obtained may not always relate directly to the acoustic signal produced by the speaker, and assessment of the acceptability of the user's speech may therefore be difficult.

The choice of a single microphone determines that the speech analysis methods have to deal with the speech waveform, since this is the only source of information available, and however, as discussed above, it contains all the information necessary for analysis of features like loudness, pitch, vowel quality, consonants.

Speech analysis methods generally consist of a number of different modules, each one addressing a single feature of the voice. There is a module for the measurement of fundamental frequency, one for the measurement of vowel quality, and so on. A detailed description of the modules used in the system described in this thesis, and of the methods used in each module is given is Section 6.3.2.3. and following. These modules are:

- real-time loudness tracker: measuring the value of the sound intensity in a vocal production
- real-time pitch tracker: measuring the value of the fundamental frequency, and therefore also able to identify which parts in a speech production are voiced or unvoiced. Some algorithms can also give information about some aspects of voice quality, such as harshness
- real-time formant tracker: measuring the value of the formants F_1, F_2, F_3... which identify the different vowels
- phonetic segmenter: which allows various measures at a segmental level.

Errors in deaf speech, discussed in Chapter 2, can influence the behaviour of some of the speech analysis modules, as follows.

Vowels, consonants, timing

In order to focus on sounds embedded in a word or sentence, an utterance has to be automatically phonetically segmented. In this case, errors in segmental production such as changes in vowel quality, and in consonants, such as voicing errors, place of articulation errors, manner of articulation errors, omission errors and consonant cluster errors, all have to be forecast and included in the model used by the segmenter. The inclusion of variations in the acoustic models used in the segmenter is necessary also in the case of errors involving suprasegmental features, such as speech rate, speech segment durations, pausing, and breath control, which may cause lengthened vowels and inappropriate pausing.

Fundamental frequency

Some speakers show excessive variations in the change of fundamental frequency (Monsen, 1979; Smith, 1975b; Stevens et al., 1978). For this reason pitch tracking algorithms should be able to work over a wide range of frequencies. Algorithms which have to be set for a specific user's voice range in order to work properly should be avoided.

In some cases there may be problems when the feedback derived by a pitch tracker is used to correct a fundamental frequency that perceptively sounds too high: Wirz (1987) observed that in some speakers the fundamental frequency seems higher, but in fact this is an effect due to laryngeal tension, which causes an increase in the spectral energy at the middle frequencies (Laver, 1980), sometimes leading to a perceived raised pitch. With these speakers, a pitch tracker alone may not be adequate to give the correct feedback, and ideally a pitch tracker should be associated with a laryngeal tension detector, However, the effect of laryngeal tension on perceived pitch is not particularly high, therefore this aspect does not generally constitute a problem in normal speech rehabilitation practice.

Voice quality, phonation

Pitch in creaky, harsh and breathy voice is generally more difficult to detect reliably by pitch tracker algorithms (Vieira, 1996). Furthermore, in the case of autonomous use of the rehabilitation system, it is necessary to signal to the users when their voice quality changes to some possibly harmful modes, such as excessive creakiness, that produces a strain in the vocal folds. Some algorithms have an intrinsic capability to measure *jitter* and *shimmer*[1], which indicate the amount of harshness and creakiness in the voice.

Loudness

There are no real problems when loudness tracking algorithms are used with the voice of deaf people, even if there may be remarkable and uncontrolled variations in level (Marhony, 1968). Nevertheless, since the goal of loudness analysis is user self-monitoring via some sort of feedback, it is necessary in some way to calibrate this "closed-loop" system, especially if close-talking microphones are used. These microphones are largely used in therapy applications for their convenience and excellent rejection of ambient noise. However they show an output level strongly dependent on the distance from the mouth. At the recommended distance, which for typical models is about one centimetre from the mouth, the microphones produce an output level higher or lower than the output level shown at a distance only a few millimetres closer or further away. This is due to the "proximity effect". This intrinsic characteristic of the microphone is generally not a problem in the case of normal recordings (the change not only in level, but also in voice quality due to the proximity effect may be also considered as a feature), but it has to be taken into consideration in the case of loudness measurement. Optionally a table microphone can be used, at a distance of about 15 to 30 cm from the mouth. At such a distance the proximity effect does not apply, and changes in distance from the microphone cause a lower change in the microphone output. It is clear that in this application an automatic gain control (AGC) device cannot be used.

[1] *Jitter*: frequency (or period) perturbation, that is the variability of the fundamental frequency or, reciprocally, of the fundamental period. Jitter measurements are concerned with short-term variation, how much a given period differs from the period that immediately follows it. It is then a measure of the frequency variability not accounted by voluntary changes in F_0.
Shimmer: amplitude perturbation. Like frequency perturbation scores, measurements of shimmer serve to quantify short-term instability of the vocal signal.

6.3.2.3 Speech analysis algorithms for hearing-impaired speech

Loudness

Sound is the sensation detected by the ear, caused by changes in air pressure. Loudness is the perceptual attribute that corresponds to the magnitude of the changes in pressure. It is influenced by the fundamental frequency and spectral properties of the stimulus (Fletcher, 1934, 1953).

The sound intensity (I) is defined as the quantity of energy in the time unit (power) through a unitary area perpendicular to the direction of propagation. The intensity is also proportional to the square of the pressure[1].

In Acoustics, the use of *sound intensity level*[2] is more common than sound intensity (I).

Another popular measurement is the *sound pressure level*[3] (SPL). The SPL is a recognised standard for quantifying the absolute sound pressure level of a sung or spoken utterance, since in actual practice sound intensities are usually measured by means of a microphone, which converts sound pressures into electrical voltages. A scheme for measuring sound pressure level is shown in Figure 6.16.

---

[1] *Intensity* is proportional to the square of the pressure, as follows:

$$I = \frac{P}{A} = \frac{p^2}{\rho v}$$

where: p = sound pressure (N/m^2)
ρ = specific air mass (constant, in Kg/m^3)
v = propagation speed (constant, in m/s)
I = sound intensity (W/m^2)
P = effective sound power (W)
A= area of sound wave (m^2)

[2] *Sound intensity level* (*IL*) is measured against a reference intensity I_o as follows:

$$IL = 10 \log \frac{I}{I_o}$$

IL is measured in decibels (dB). The reference intensity I_o is $I_o = 10^{12}$ W/m^2.

[3] The SPL is the logarithmic transform of the ratio of two sound pressures:

$$dB_{SPL} = 20 \log_{10} \frac{P_1}{P_r}$$

The reference pressure is, by international convention, expressed as 20 micropascals (1 Pascal is 1 newton per square metre).

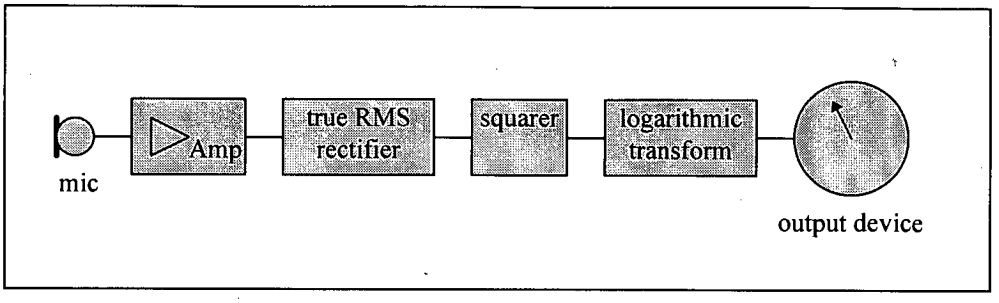


Figure 6.16. Block diagram of sound pressure level measurement

The microphone transduces acoustic pressure into voltage. The amplifier increases the very low voltage (a few millivolts) from the microphone to a more usable voltage level. The root-mean-square (RMS) voltage is calculated in order to obtain the power of the sound wave (this is accomplished in analogue implementations by a circuit whose output is proportional to the integral of the voltage). The squarer multiplies the result of the RMS block by itself, since the power is proportional to the square of the pressure (or the square of the voltage). The last block computes the logarithm.

Note that measuring SPL requires a system where the microphone and the other components of the metering system should be calibrated with some precise sound source, and never readjusted without a subsequent recalibration. It was already noted that mouth-microphone distance can greatly affect the measure (the intensity of the sound is approximately inversely proportional to the square of the distance from the source).

However, in order to give an effective visual feedback of voice level, an accurate SPL analysis is not required, for the following reasons:

1. The speech feature of interest is the sound level in relative terms rather than in absolute terms. The goal of the feedback is to show if the sound level of an utterance is lower or higher than a desired "optimum" whose level is chosen case by case by the therapist.
2. The visual feedback rather than showing the SPL value as a number will be implemented using some sort of animation. The display has to make sense in "visual" terms.
3. As mentioned before, loudness is not easily correlated with sound pressure, and is influenced by the fundamental frequency and spectral properties of the stimulus.

The first point removes the necessity for a calibrated system, with evident advantages in terms of simplicity of use, and possible cost of the system. The second and the third points allow the use of different methods for measuring speech levels, and selection of the one which gives the best results in visual terms.

Implementation

A typical session of speech rehabilitation therapy supported by visual feedback showing loudness involves the use of a continuous vowels, for example long /a/ or /e/. Sometimes repeated CV sequences, such as "ba ba ba" are used. For a practical implementation of the visual feedback, there is no need to calculate the true RMS of the waveform, and an average of the absolute value of speech frames having a duration of 50-100 ms. is adequate. As mentioned before, intensity-to-loudness conversion is not straightforward. However a simple approximation does exist, stating that the perceived loudness, $L(\omega)$, is approximately the cube root of the intensity, $I(\omega)$:

$$L(\omega) = I(\omega)^{\frac{1}{3}}$$

This is not the case for very loud or very quiet sounds, but it is a reasonable approximation for speech (Robinson, 1996). In the case of connected speech, pauses that are either physiologically or linguistically required in an utterance were taken into account, since they reduce the overall average intensity. Furthermore, depending on the particular goal that the therapist is trying to achieve, a fast or smoothed response of the visual feedback should be chosen. This was achieved through a simple low-pass smoothing filter. Figure 6-17 shows the block diagram of the steps performed to analyse the speech level.

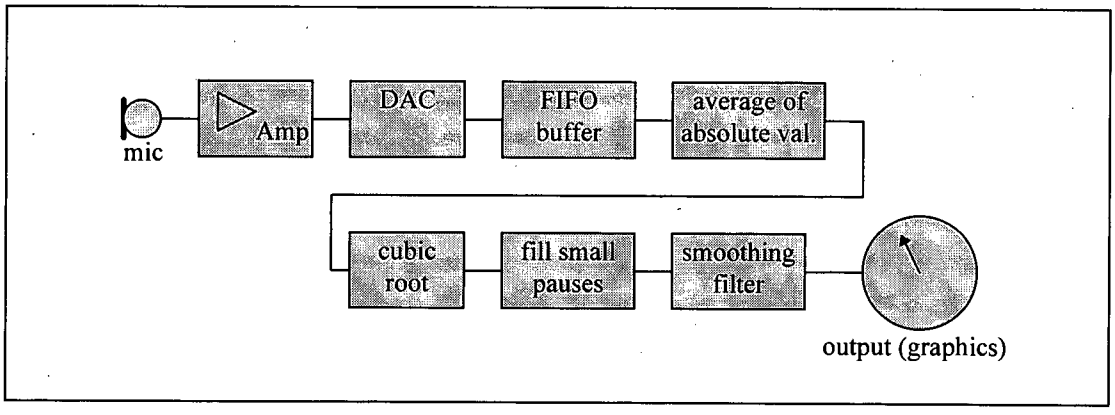


Figure 6-17. Speech level analysis

Small pauses were filled by a simple method, which works approximately as a "peak VU-meter indicator" as found on audio equipment. If the level of a 100 ms. frame is below a threshold (of silence), the value of the level of the previous frame is repeated in the output for a maximum of two frames. Optionally this "pause filler" can be disabled.

The smoothing filter is implemented as a single tap infinite impulse response (IIR) filter, as shown in Figure 6.18. The filter response is easily set by the therapist.

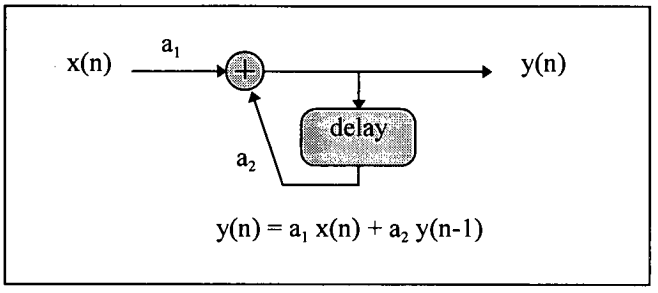


Figure 6.18. Smoothing filter

The practical difference between such an IIR filter and a finite impulse response (FIR) filter with a similar response is that the IIR gives a slightly more natural "damping" of the visual feedback (caused by its infinite response). Furthermore the IIR implementation is computationally more efficient.

After being smoothed, the result of the speech level analysis is ready to be fed as an input to the graphics section of the visual feedback, after being scaled and non-linearly transformed in order to achieve the desired visual result.

As seen above, elaborate signal processing is not necessary for measuring the speech level. The computing power saved by calculating the average instead of the RMS can be used for other purposes, such as further speech processing and graphics.

Fundamental frequency

The speech fundamental frequency, generally indicated with F_0, is the rate at which the vocal folds vibrate per unit time. It is measured in Hertz (Hz) which are equal to cycles per second. Speech fundamental frequency detectors are usually called pitch detectors, or pitch determinators, even if these terms are not exactly accurate, since pitch is non-linearly related to fundamental frequency, and influenced by other factors, such as sound level and spectral properties (Goldstein, 1973; Fletcher, 1934; Laver, 1980). Anyway following the convention universally adopted, these algorithms are called here "pitch detectors", dealing with the difference between fundamental frequency and pitch later in this Chapter.

The task of a pitch detector is not simple for a number of reasons, including non-stationarity of speech, characterised by a constantly changing spectrum caused by sudden changes in the position of the vocal tract; narrow-band formants at low harmonics; non-uniform intensity; variations between successive periods, particularly noticeable in creaky voice; and wide frequency range, spanning from 50 to 800 Hz for an unknown speaker (Hess, 1983). The algorithms employed for pitch detection cover a range of techniques which use both time domain and frequency domain representations of

speech, as well as hybrids using linear prediction coding (LPC) and autocorrelation. Time domain methods try to detect the periodic peaks in the waveform using various logical processes. Among the frequency domain methods, two popular ones are the *cepstrum method*, which removes the vocal tract effects by separating spectral envelope and fine structure by inverse Fourier transform of the log-power spectrum, and the *period histogram method* which finds the fundamental frequency as the common divisor of high harmonic components. Hybrid methods use autocorrelation of the waveform, or autocorrelation of the residual signal of LPC analysis to find the pitch.

Selecting a pitch detector algorithm

In choosing the best algorithm for speech rehabilitation the following criteria must be considered:

- Accuracy in estimating the fundamental frequency
- Accuracy in making a voiced-unvoiced decision
- Suitability of the algorithm for the chosen platform
- Speed of operation
- Capability of the algorithm to give information on voice quality

Some comparative evaluations were considered where a variety of pitch detector algorithms were investigated:[1]

1. Cepstrum pitch determination (CPD)
2. Feature-based pitch tracker (FBPT)
3. Harmonic product spectrum (HPS)
4. Integrated pitch tracking algorithm (IPTA)
5. Parallel processing method (PP)
6. Super resolution pitch determinator (SRPD)
7. Enhanced version of SRPD (eSRPD)
8. K. Schafer-Vincent (KSV)
9. Modified K. Schafer-Vincent (mKSV)

The first seven algorithms were evaluated (Bagshaw et al., 1993), using a database containing approximately five minutes of speech, read by one male and one female, biased towards utterances containing voiced fricatives, nasals, liquids and glides, which are generally difficult to analyse, and comparing the results with reference contours from post-processed laryngograph data. The goal was to find the most suitable algorithm for intonation analysis (Rooney et al., 1992). The chosen criteria were: 1) medium/high accuracy in estimating the fundamental frequency, with a low instance rate of

[1] references: CPD: Noll, 1970; FBPT: Phillips, 1985; HPS: Schroeder, 1968 and Noll, 1970; IPTA: Secrest & Doddington, 1983; PP: Gold & Rabiner, 1969; SRPD: Medan et. al., 1991; eSRPD: Bagshaw, 1993; KSV: Schafer-Vincent, 1983; mKSV: Vieira, 1996.

pitch doubling/halving (octave errors), 2) good accuracy in making voiced-unvoiced decisions, 3) consistent performance between male and female speech. The most suitable algorithms were the SRPD (Medan et al., 1991) and its enhanced version eSRPD (Bagshaw at al., 1993). The latter features a post processing algorithm expecially designed for compensating the intonation slope that prevents a correct visual evaluation of intonation in long sentences. Speed of execution was not considered a particularly important issue for the application required, and it was not taken into account in the comparison.

The last two algorithms were compared (Vieira et al., 1996) against the SRPD using a database containing 4.5 minutes of speech read by fifteen adult subjects (eight females, seven males) and simultaneously recorded on the two tracks of a DAT recorder using a head-mounted microphone and an electroglottographic device (Laryngograph) calibrated in order to reduce errors due to EGG baseline fluctuations (Vieira et al., 1995). The comparison showed that the modified K. Schafer-Vincent (mKSV) method significantly reduces the occurrence of high errors (above +10%) in respect of the other methods, and there were small improvements in all other parameters, with the exception of voiced-to-unvoiced errors. Furthermore the speed of the mKSV method was remarkably 8.3 times faster than that of the eSRPD method[1]. For these characteristics the mKSV was chosen as the best one for our application. Table 6.1 shows the results of the comparison.

	H	L	|F|	F SD	U2V	V2U
Females						
SRPD	2.53	2.46	1.15	1.41	1.87	6.41
KSV	0.43	0.80	1.97	1.91	0.81	7.64
mKSV	0.26	0.62	1.93	1.93	0.62	8.19
Males						
SRPD	2.28	3.49	1.32	1.46	1.56	9.56
KSV	0.52	1.03	2.14	1.99	0.69	10.85
mKSV	0.31	1.01	2.05	1.86	0.54	11.41

Table 6.1. Comparison between 3 methods. All values are expressed in percentages. H (high errors), i.e. above +10%, and L (low errors), i.e. below -10%, are proportions of the number of F_0 estimates in the contours being assessed; |F| is the absolute mean of the fine errors (i.e. within ±10%) and F SD is the respective standard deviation; unvoiced (or silent)-to-voiced errors (U2V) and voiced-to-unvoiced (or silent) errors (V2U) are shown proportionally to the utterance duration (from Vieira, 1996).

[1] C code on a 486 33 MHz machine.

Schafer-Vincent algorithm

Algorithm KSV (Schafer-Vincent, 1983) works on the time-amplitude representation of the speech signal. It detects quasi-periodic parts of the signal, first determining potential "period-twins" and then deciding whether to validate these period-twins as components of a "period-chain". A simplified explanation of the algorithm is given here, leaving details to the original paper. At first the algorithm identifies "significant points" in the waveform, choosing no more than one maximum and one minimum every 2 ms. Every new significant point is then tested in order to be combined with previously-found significant points, attempting to identify a set of three significant points which characterise two periods in the speech signal (period-twins). In order to be part of a period-twin, the two periods have to satisfy the following characteristics: to have a period corresponding to a frequency above 50 Hz, similar in period duration with a variation (jitter) below 10%, above a noise threshold limit, exhibiting the highest amplitude value in the period-twin, having a cycle-to-cycle amplitude variation (shimmer) below 50%, and having a similar waveform shape. All the period twins matching these constraints are stored in a period-buffer and are concatenated together if they satisfy another set of rules taking into account the similarity between period-twins and their time-alignment. Several "chains" are built in parallel, and more than one may be validated at the same time for generating an output. The output values are time-stamped and stored in an output buffer that may then contain overlapped results (which happens often since normally both maxima and minima can be used as period boundaries). When this happens an average of the overlapping values is calculated. Also a chain may be validated and stored in the buffer marked as a substitute for a previously validated one, still in the output queue. In this case the buffer provides a correction mechanism.

The modified algorithm (Vieira, 1996) speeds-up some of the tests without significant change in accuracy. The modification consists mainly of two points: when trying to identify a period-twin the search was limited at the first successful attempt, instead of considering all possible period-twins. The second improvement in speed was achieved by rewriting the test that compares the waveform shape, using a simpler but equally effective method. The two modifications together resulted in an improvement in computational speed of 30%.

Implementation of the real-time version of the pitch tracker

The algorithm was initially implemented on a DSP32C board, and written in C, with some parts in assembler (samples acquisition, anti-aliasing and buffering). This simplified the testing of the algorithm in real-time, making it independent from the PC's processor, other tasks and the computationally complex Windows multimedia audio drivers. On the 50 MHz DSP the algorithm used from 70% (in case of silence) to 130% (in case of high pitch) of the processor time[1]. The varying load depends on the increasing number of tests that the period-twin has to pass in order to be considered valid. Silent frames fail the first test immediately and do not need further processing. High pitched speech needs more processing than low pitched speech because of the higher number of period-twins and relative tests they have to pass.

When eventually the algorithm was ported to the main processor of the PC, (a Pentium 90 MHz) the speed of execution was faster than for the DSP implementation, using from 25% to 45% of the processor time, depending on the signal input. It may be surprising to achieve a better performance in a signal processing task from a general purpose CPU than from a specialised DSP, but this is can be explained by two facts: 1) the same C source generates a longer binary code on the DSP32C than on the Pentium, therefore requiring more clock cycles for execution; 2) the mKSV algorithm does not take advantage of the MUL (multiply) or MAC (multiply and accumulate) instruction that make a DSP so efficient for most signal processing algorithms. Being a time-domain type of algorithm, most operations are comparisons and movements of data in memory, where a general purpose CPU gives its best.

Porting the fundamental frequency detector on the PC's CPU caused some problems. The operating system used in the first prototype (Windows 3.1) is not a pre-emptive OS, so a process can take the CPU time for all the time it needs. This causes the many processes running simultaneously to stop and start in an irregular way. In the specific case, there was the possibility that some speech samples could be lost. Furthermore, the graphic feedback which was designed in order to fit in the remaining CPU time, was obviously using this computational power in a highly variable way, requiring more or less CPU attention in relation with the changes of graphics caused by the changes in the pitch of the

[1] The CPU performance was monitored in the following way: every time the algorithm finished processing a frame, the code jumped to execute a loop whose number of clock cycles to execute was known. The task of this loop was to increment a variable in a memory location shared between the DSP and the CPU of the PC. The variable was read periodically by the CPU of the PC, and reset after each read. The value showed how many cycles of "free time" the DSP had spent, making it possible to calculate the percentage of DSP load. When this value dropped to zero, because the loop had no time left to run, the algorithm was not working in real-time any more. The higher-priority data aquisition was then queuing new samples in an opportunely dimensioned data buffer with a rate higher than the algorithm could process. When the speech input was removed, the algorithm started processing samples at a rate higher than that in which the new (silent) samples were queued. When the variable counting the "idle" loops was starting to be incremented again, the algorithm was running in real-time. So it was possible to measure the CPU load by computing the time it took to reach the real-time regime again. In this context 130% CPU load means 1.3 times real-time.

speaker's voice. This caused a possible overload of the CPU, with the result that sometimes the pitch tracker was not running in real-time for some short periods. The effect was a noticeable delay in the response of the graphic display, giving a confusing feedback to the user.

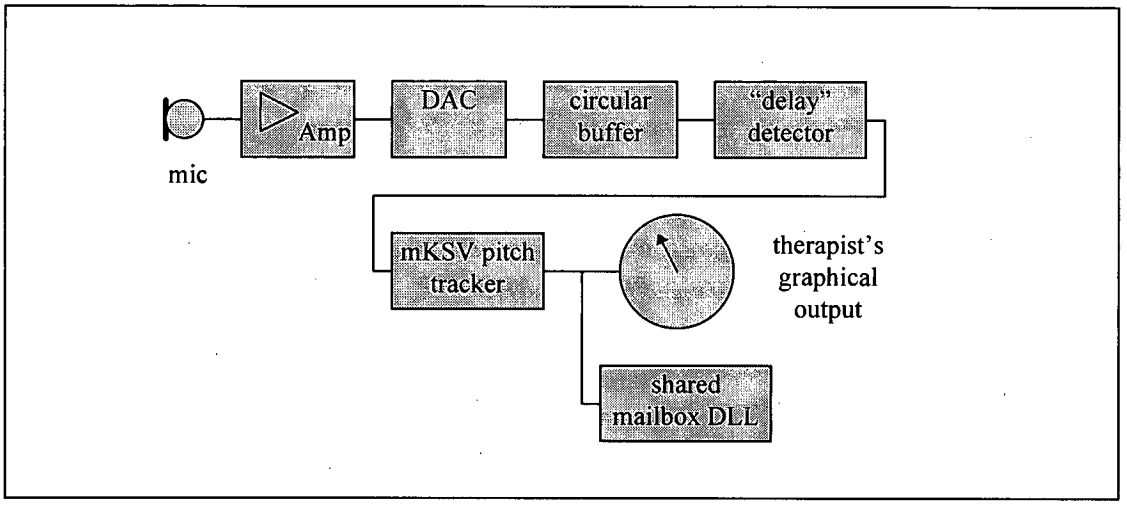


Figure 6.19. Real time pitch tracker implementation

A simple mechanism was implemented to avoid this delay, triggered by the detection that the pitch determination was noticeably late with the incoming speech samples. In such a case, the input FIFO buffer was flushed as necessary in order to re-sync the pitch tracker, taking care to smooth the possibly inaccurate results from the pitch tracker caused by the "hole" in the input data (see

). The mKSV algorithm behaved very well in this unusual situation, because of its built-in capability to re-sync chains[1]. Since this real-time version of the pitch tracker is meant to be used with continuous vowels, there is no negative effect for the possible lack of analysis in small sections of the input data. For more detailed intonation analysis a non-real-time version was implemented, as discussed later in this section. Another modification made was the option to set the pitch tracker to look for significant points in a running window shorter than 2 ms., so increasing the maximum measurable pitch from 500 Hz to about 800 Hz, necessary for use with children's speech.

An interesting capability of the algorithm is the measurement of fast changes in pitch period (jitter), which is useful information for detecting potentially dangerous situations for the vocal folds, such as creaky voice. This capability can be exploited to alert the user when their voice becomes harsh or creaky, a common case for hearing-impaired people trying to control a visual "drill" with their voice.

The pitch tracker was written entirely in C using Microsoft Visual C++, and was realised as an independent programme having a "therapist's" graphical output (see Figure 6.20) consisting of a real-time waveform window (top left) and a scrolling pitch window, resembling a medical pen recorder on paper (main window).

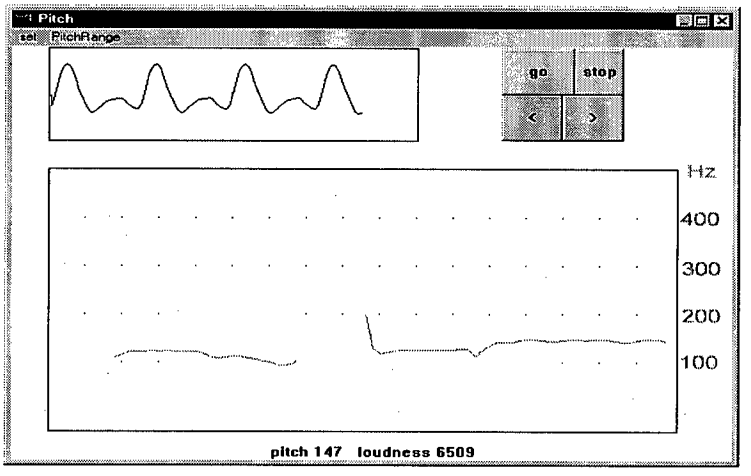


Figure 6.20: *Therapist's* graphical output of the real-time pitch tracker

Voiced / unvoiced sections of speech were shown by interrupting the pitch plot in the graphic in the unvoiced sections.

The pitch tracker was interfaced to the control programme and its graphic displays by means of a DLL (Dynamic Link Library) which was implementing a shared mail-box between the two modules, and exchanging commands and results between them.

Implementation of the non-real-time version of the pitch tracker

Another version of the mKSV pitch tracker was implemented for use with a non-real-time display, giving delayed feedback for an intonation rehabilitation module. Such a module requires a higher temporal resolution than the real-time modules, since it is meant to be used with short utterances such as words or short phrases, instead of with continuous vowels.

[1] Chains may be built by linking chain elements resulting from the detection of positive significative peaks in the waveform, and chain elements resulting from the detection of negative peaks as well, and they are mutually out-of phase between them.

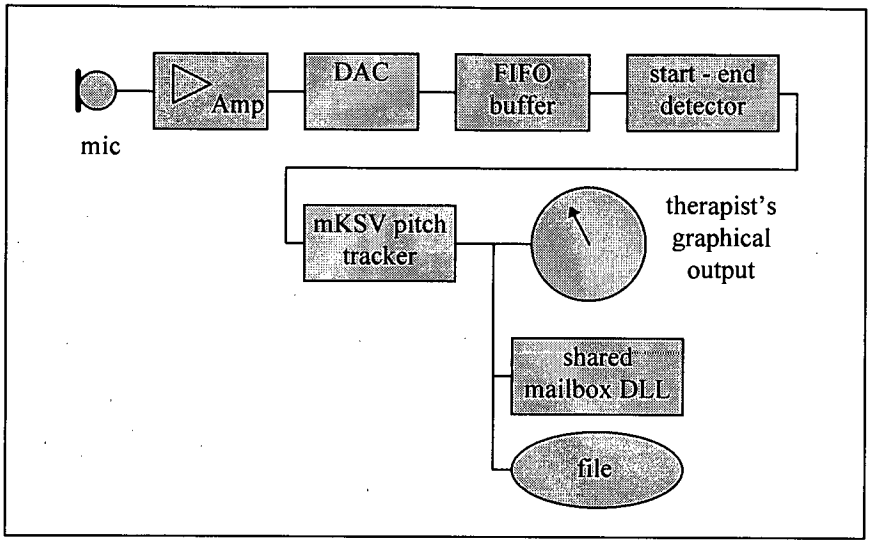


Figure 6.21. Non-real-time pitch tracker implementation

The module was implemented as follows (see Figure 6.21): the data samples from the "low level audio services" of the Windows operating system came in buffers of about 100 ms. of speech. The buffers were analysed by an energy-based start-end detector, and also stored in FIFO buffer capable of containing about 10 seconds of speech. At the detection of the end-point of the utterance, the data acquisition was stopped and the mKSV pitch estimator was started. The results were written in a temporary file, for easing further operations by external modules, and for giving the possibility to store pitch data on disk. Simple graphics showing the pitch track were implemented for debugging and calibrating purposes. The exchange of commands and results with the controlling module was done via a DLL used as a shared mailbox.

The mechanism for re-synching the algorithm in case of CPU overload was not implemented because it was not necessary. The response time was between about 0.2 times and 0.35 times real-time[1] (meaning that for example one utterance 1 second long took from 0.2 to 0.35 seconds to be analysed, once the end-point was detected). The increased speed in comparison with the real-time implementation may be explained in this way: in the real-time implementation there was an overhead due to the data acquisition. Furthermore the non-real-time implementation possibly took advantage of a much higher number of "cache hits" because the data acquisition and the graphic module were not interrupting the flow of operation of the pitch tracker module (small enough to be entirely cached).

[1] On a Pentium 90Mhz.

Vowels

A powerful description of the acoustic characteristics of vowels is called "the source-filter theory of vowel production" and was discussed in Chapter 2. To summarise, in this theory, the glottis is the source of signals, rich in harmonics, which are then filtered by resonances in the vocal tract. The harmonics in the source whose frequencies are near the resonance peaks of particular vocal tract configuration are enhanced, the others are attenuated. The resonance peaks of the vocal tract are called formants. The perception of vocalic sounds depends mostly on the formant characteristics that are reflected in the sound wave radiated by the lips. These formant characteristics depend on the shape of the vocal tract, determined by the position and shape of the tongue, jaw opening, etc. Changes in vocal tract shape change its resonances, and different vowels are perceived. The vowel perceived is not influenced significantly by the glottal source. However, the spectral characteristics of the glottal source interact with the resonant properties of the vocal tract to produce the spectral

Non parametric methods:	
Short-time autocorrelation	Spectral envelope and fine structure are convoluted. The simple, easy algorithm facilitates hardware realisation.
Short-time spectrum	Spectral envelope and fine structure are multiplied. This fast algorithm can be realised by FFT.
Cepstrum	Spectral envelope and fine structure can be separated in *quefrency* domain. Two FFTs and log are necessary.
Band-pass filter bank	A global envelope can be obtained. This is appropriate for real-time processing.
Zero-crossing analysis	Formant frequencies can be obtained by combination with a band-pass filter bank. This can be realised by simple hardware.
Parametric methods:	
Analysis-by-synthesis	Precise modelling is possible. Accurate formant frequencies can be obtained. Complicated iteration is necessary.
Linear Prediction Coding	Simple all-pole spectrum modelling. Parameters can be estimated from autocorrelation or covariance without iteration.

Table 6.2. Major methods for analysing speech spectra and their principal features (from Furui, 89).

characteristics of the radiated sound wave. Therefore, analysing the sound wave in order to identify a vowel requires a separation of the vocal tract filter characteristics from the glottal source characteristics. In a short-time speech spectrum this means separating the envelope from its fine structure. Methods for envelope extraction are categorised as parametric and non-parametric. Parametric methods take advantage of our understanding of how the speech signal is produced. A suitable model of the human speech production at the vocal tract level is selected and the parameters representing the model are adjusted in order to fit the signal. Non parametric methods do not model the signal, so they can be generally applied to various signals. The most common methods for analysing speech spectra are shown briefly in Table 6.2. If the model fits the signal well, parametric methods can represent the features of the signal more effectively than non-parametric methods. Figure 6.22 shows a comparison between the Cepstrum method and the Linear Prediction Coding (LPC) method. The LPC envelope clearly tends to follow the spectral peaks more strictly than the spectral envelope obtained through the FFT cepstrum[1].

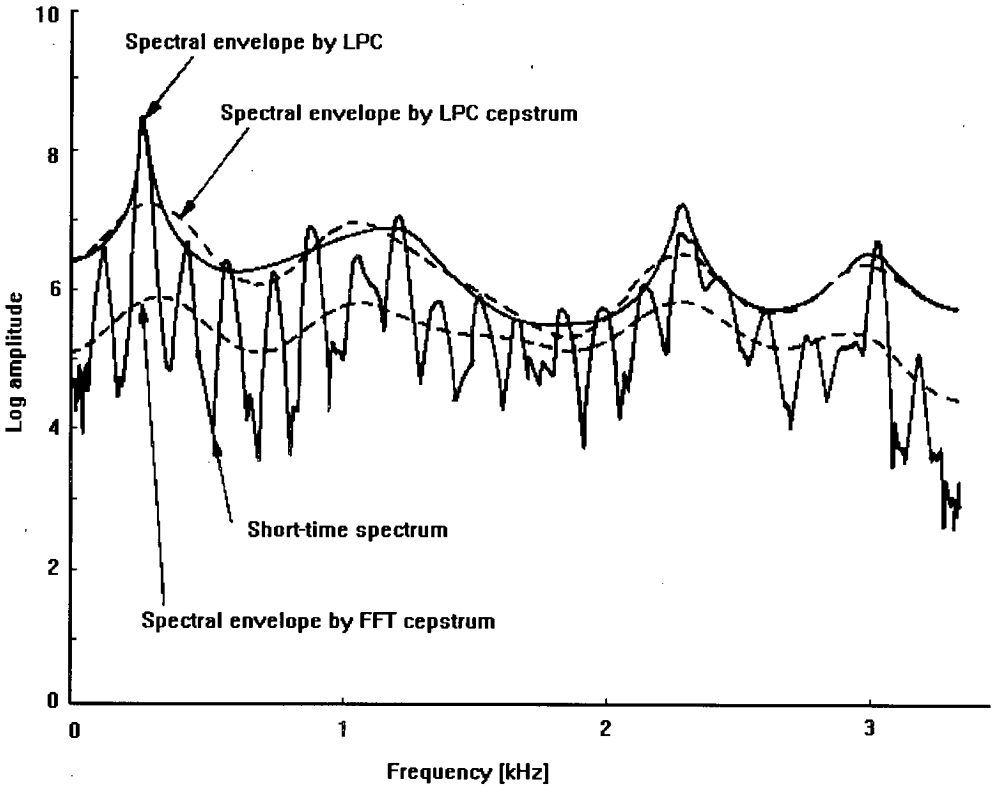


Figure 6.22. Comparison of spectral envelopes by LPC, LPC cepstrum and FFT cepstrum methods (from Furui, 1989)

[1] Figure 6.22 shows also the envelope obtain with LPC cepstrum, a cepstrum derived through the LPC model. The envelope results in something in between the FFT cepstrum and the LPC envelopes.

The LPC model provides an easy and effective way to separate the spectral envelope from the fine structure, and comparison with spectra obtained by other means show that formants derived by this method are accurate about 85-90% of the time without further processing, and involve less computational load (Fallside & Woods, 1985).

Using this model[1], the speech wave and spectrum characteristics can be efficiently represented by a small number of parameters, obtained by relatively simple calculations. These parameters can be used to extract the formant frequencies, which characterise the different vowels, by root solving[2] or by extracting the short term envelope[3] (as in Figure 6.22) and then by peak-picking to establish the formants (see Figure 6.23). The latter technique is computationally much simpler. Both methods produce raw data requiring further processing in order to closely follow changes in the vocal tract.

---

[1] For details about LPC see for example Markel & Gray, 1976. A brief description of the method follows: a sampled speech waveform $s(n)$ can be approximated by another one $s'(n)$ by linearly predicting from the past p samples of $s(n)$:

$$s'(n) = \sum_{k=1}^{p} a_k s(n-k)$$

The parameters a_k can be determined by minimizing the mean squared difference E between $s(n)$ and $s'(n)$ over N samples of $s(n)$:

$$E = \frac{1}{N} \sum_{n=0}^{N-1} [s(n) - s'(n)]^2$$

To minimize E, two methods that give approximately the same results are generally used, the *autocorrelation method* and the *covariance method*. The first one (using Durbin's recursive solution method) needs a smaller number of calculations than the second one (using the Cholesky decomposition method), and assuming p=10 is about three times more efficient, and assured to be stable, but needs several pitch periods to be contained in the analysis window for meaningful results with voiced speech (Markel & Gray 1976).

[2] Finding the poles of the vocal tract transfer function by solving the roots of the polynomial equation

$$\sum_{k=1}^{p} a_k z^{-k} = 1$$

a_1 .. a_p are the LPC coefficients. The p roots z_k of the function can be found with a root-solving procedure. From the roots it is possible to obtain the frequencies and the bandwidths of the formants. Since the roots occur in complex conjugate pairs (as the polynomial is real), the angle defines the formant frequency, and the distance from the unit circle defines the bandwidth.

[3] As the power spectrum of the impulse response of the filter $H(z)$, represented by the coefficients a_k, at N equally spaced samples along the unit circle in the z-plane:

$$H(z) = \frac{a_0}{1 - \sum_{k=1}^{p} a_k z^{-k}}$$

where

$$z = \exp[j(\frac{2\Pi n}{N})]$$

for $n = 0,1,\ldots,N-1$. N can be chosen arbitrarily large for increasing frequency resolution. With N=256 the frequency resolution is about 40 Hz, assuming a sampling rate of 10kHz.

The raw data may contain false peaks, showing formants where there are none, or merged peaks, which hide formants. The methods for solving these problems generally impose constraints in continuity and rate of change of formant tracks. Markel (1972) presented a simple decision-making algorithm for automatically extracting the first three formant trajectories from the raw spectral data. This will be briefly explained, and followed by two more complex algorithms that give increased accuracy.

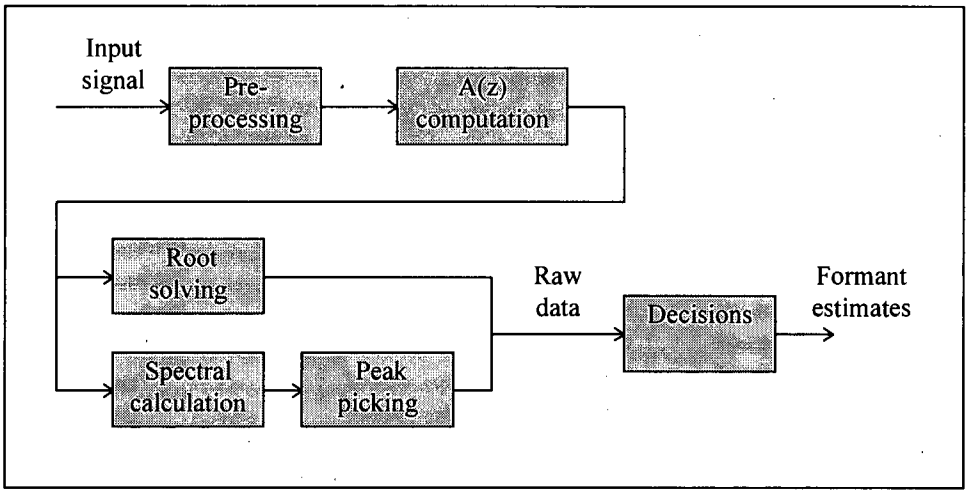


Figure 6.23. Formant trajectory estimation (from Markel & Gray, 1976)

Algorithm 1

Markel's algorithm counts the peaks in the spectrum below 3 kHz that satisfy a bandwidth threshold such as <500 Hz. By choosing an adequate number of LPC parameters, there should never be more than four peaks that satisfy the criteria. Depending on the number of peaks detected, different strategies are used. If there are three peaks (approximately 85-90% of the time) these peaks are assumed to be the formants. If only one peak satisfies the criteria (less than 1% of the time), that peak is assigned to the formant in the previous frame with the closest value. The other two values are copied from the previous frame. If there are two or four peaks, a case of omission or insertion, respectively, has to be dealt with. In the case of two peaks, the two slots are filled following similarity with the previous frame. The third slot is copied from the previous frame's corresponding formant value. If there are four peaks, three out of four values are assigned to the formant slots having the closest values in the previous frame. As initial conditions the algorithm assumes the neutral vowel frequencies of 0.5, 1.5 and 2.5 kHz for the three formant slots. This algorithm is accurate for male voice, since the third formant is assumed below 3 kHz. For female or child voices, where the third formant can be greater than 3 kHz, the algorithm does not perform well, since when 3 peaks below 3 kHz are detected they are directly assigned to the first three formants.

Algorithm 2

Markel's method relies strongly on continuity of formants. According to McCandless (1974) this may be dangerous for two reasons: 1) formant frequencies can change quickly, for example in less than 5 ms. at the boundary between a nasal and a vowel; 2) if a frame was wrongly analysed, for example if an unvoiced frame was considered voiced, and an attempt made to detect its formants, it may cause errors in many of the following frames. In order to solve this problems, the McCandless method uses the following approach: 1) it tries to find a reliable starting point in order to select formants from a stable vocalised section of speech, and then moves in both directions to trace adjacent areas; 2) it uses continuity only when it gives reasonable results, and relies on other techniques when this fails. Another characteristic of the method is that the allocation of peaks in the formant slots is carried-out in parallel instead of in series as with other methods. For example, each peak can be a candidate for more than one formant slot at the same time. The final decision is taken considering all the candidates in parallel. In order to find a stable vowel from which to start, a pitch tracker combined with an analysis of the energy in the band between 640 and 2900 Hz is compared with the total energy in the spectrum. This selects the centre of vocalised areas. The formant tracking proceeds until it reaches the boundaries of a previously processed area, or until an unvoiced frame is encountered. Then it jumps to the centre of the next unprocessed stable vowel.

The processing of each frame is accomplished by a six-step task whose goal is to fill four formant slots with peaks, based on estimated frequencies, trying to recognise spurious or missing peaks. Spurious peaks are deleted by a decision procedure. Dealing with missing peaks is more complex. When they cannot be recovered using continuity constraints, an attempt to separate two merged formants is done by recalculating the spectrum with a formant enhancement technique. This technique, analogue to the chirp-z transform (Rabiner et al., 1969) consists of re-computing the spectrum on a circle of radius less than 1. Since the contour comes closer to the poles, their peaks are enhanced, and hopefully a couple of merged peaks are separated.

Algorithm 3

Possibly the best results currently achievable use a dynamic programming approach to speech parameter estimation (Ney, 1983). As with the McCandless method, 1) a number of possible formant frequency candidates is generated, and 2) a subset of these candidates are labelled as formants. The source of candidates is generated in the original algorithm by LPC analysis followed by a complex-root solution, producing a set of frequencies and relative bandwidths, but any other source of candidates can be used instead of the LPC roots. All the possible mappings of these candidates for N formants are considered for every frame analysed. The mappings are then used as nodes in a Viterbi decoder lattice. Each formant mapping is associated with a local cost calculated by the reasonableness of the formant frequencies and formant bandwidths. The connection of a node in the current frame with a node in the previous frame adds a cost based on the amount in frequency change, and the

stability of the signal at the time considered. Dynamic programming gives the least cost path to go through the candidate-to-formant lattice, so finding the best set of formant trajectories. This method is currently used in some non-real-time speech analysis software[1].

Selecting a formant tracking algorithm

The McCandless algorithm is the most suitable choice for a real-time implementation without the help of extra hardware, because of its high reliability and reasonable computational load.

Implementation

The first version of the formant tracker was implemented partly on the DSP32C board, and partly on the PC processor. The DSP board was used for the following tasks: 1) analogue to digital conversion, 16 bits, 40 kHz; 2) anti-aliasing filter at four times oversampling, and conversion at 10 kHz; 3) LPC analysis (14 coefficients) window size 25 ms, shift 4 ms.; 4) data storage into a 5 sec. FIFO buffer. The PC processor (a Pentium 90 MHz) was used for the following tasks: 1) calculation of LPC spectra from the LPC coefficients; 2) McCandless formant tracker method, modified for real-time; 3) vowel normalisation (explained later in this section). In a later version of the software, the DSP board was not used, since the A/D conversion was done by a standard Windows compatible 16 bit sound card, using the "low level audio services" of the operating system, and all the speech processing could be run on the PC processor after optimisation of the slowest modules. The McCandless algorithm had to be modified for real-time use as follows: the original method looks for the middle region of a stable vowel, and then starts analysing away from that point in both directions. Since buffering the data and finding a stable region could cause unacceptable delays in the visual feedback, the initial stable region was selected using an energy based approach, which was possible since the real-time visual feedback application would typically be using continuous vowels, or repeated CV sequences. At the beginning of every new high energy section of speech, probable candidates for the first three formants were reset to fixed "neutral" values, 0.5, 1.5 and 2.5 kHz for males, and scaled higher by a factor of 1.25% for females and children. The rest of the algorithm follows the original method, with the exception of the use of spectral enhancement. Spectral enhancement was needed in case two formants were merged into a single peak, and required a new spectral analysis of the frame containing them. It was noted that the occurrences in which this might happen were very low with long vowels or repeated CV sequences (<0.2%, with the exception of some nasalized vowels, which required the spectral enhancement more often), it was therefore decided to repeat the formants detected in the previous frame instead of calling this time-consuming procedure.

[1] Xwaves+ (University of California, 1995)

Vowel normalisation

Vowels can be described by the frequency values of the first three or four formants, but what is perceived as the same vowel from different speakers may have quite different values. Several vowel normalisation techniques have been proposed, attempting to capture an invariant description of vowels, in order to reduce the variability between speakers, and possibly between male and female speech. Miller et al. (1980) describe a method which computes the ratio of the log of the formant frequencies. Bladon et al. (1984) use many features, such as spectral tilt, formant amplitude and formant bandwidth for normalising vowels via *template matching*, where the vowel spectrum is shifted on a perceptually converted plane (*bark* versus *sone*) and matched with a standard. Sydral and Gopal (1986) found that the *bark-difference* normalisation was the most effective in reducing the between-speaker variability, but yet preserving some differences between speaker groups which may reflect linguistically valid dialectal differences. In this method, vowel normalisation is obtained by considering the normalised formant values F_1-F_0, F_2-F_1, F_3-F_2. In particular, the (F_1-F_0) dimension is associated with the feature high-low, and the (F_3-F_2) or (F_2-F_1) with the feature front-back. All formant values should be expressed in bark units, according to the following transformation:

$$B = 13\arctan(0.76f) + 3.5\arctan(\frac{f}{7.5})^2$$

where B is the critical band value[1] in bark, and f is the frequency in kHz. At frequencies below 250 a low-frequency correction[2] proposed by Traunmuller (1981) is applied. The vowel normalisation was implemented using the two bark-converted couples (F_1-F_0) and (F_2-F_1), where F_0 was an estimate of

[1] In the transformation from a physical frequency measure to an auditory scale, the *critical band* scale is very appropriate for the representation of the complex speech spectrum. The critical band scale has been determined using a wide variety of psychoacoustical experiments, including loudness summation, narrow-band masking, two-tone masking, threshold of complex sounds, phase sensivity, musical consonance, and discrimination of partials in a complex tone (Scharf, 1970). A current functional view of the auditory system is that it is composed of a series of internal bandpass filters with ovelapping bandwidths (Plomp, 1975; Schroeder et al., 1979). The bandwidth of each one of these internal filters corresponds to a critical band. The critical band scale increases linearly with frequency up to 500 Hz, and then continues approximately logarithmically. Zwicker (1961) proposed that an empirically defined critical band scale be adopted as a standard tonality scale. His proposed scale divides the human auditory range below 16 kHz into 24 critical band units, or barks, named after Barkhausen, the creator of the unit of loudness level.

[2] In low-frequency correction, all frequencies below 150 Hz are raised to 150 Hz, and for frequencies between 150 and 200 Hz the corrected frequency is defined as:

$$f_c = f - 0.2\,(f - 150)$$

while for frequencies between 200 Hz and 250 Hz is:

$$f_c = f - 0.2\,(250 - f)$$

where f_c is the corrected frequency and f is the frequency in Hz.

the mean fundamental frequency of the speaker (this value was decided at the beginning of the speech rehabilitation session and set by the speech therapist).

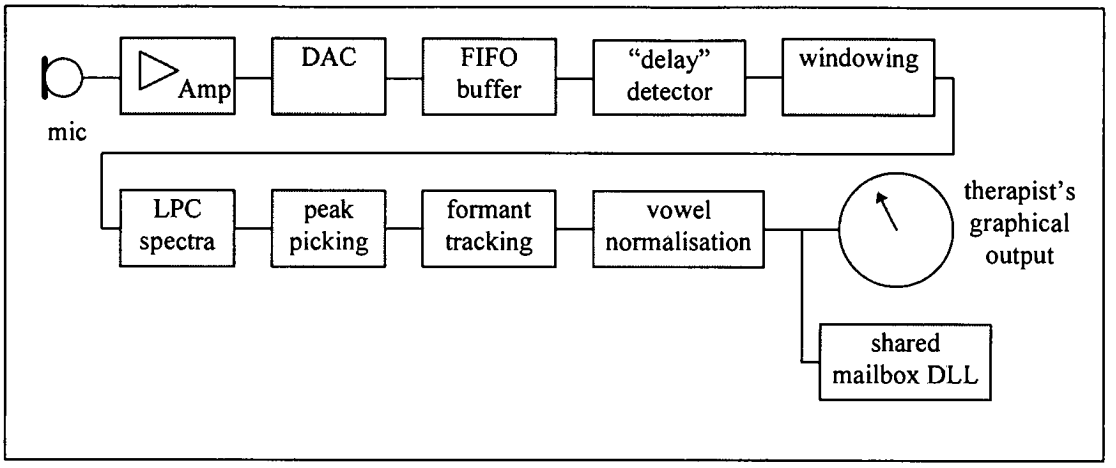


Figure 6.24. Real time formant tracker implementation

The complete real-time formant analysis, with vowel normalisation (see Figure 6.24), takes about 80% of the total CPU time on a Pentium 90 MHz. However, the same "re-synching" mechanism implemented for the real-time pitch tracker in case of CPU overload was implemented here, allowing CPU intensive graphics to run concurrently without delaying the visual feedback, and also to make this algorithm usable even on slower CPU's. Figure 6.25 shows the *technical* graphics of the real-time formant tracker.

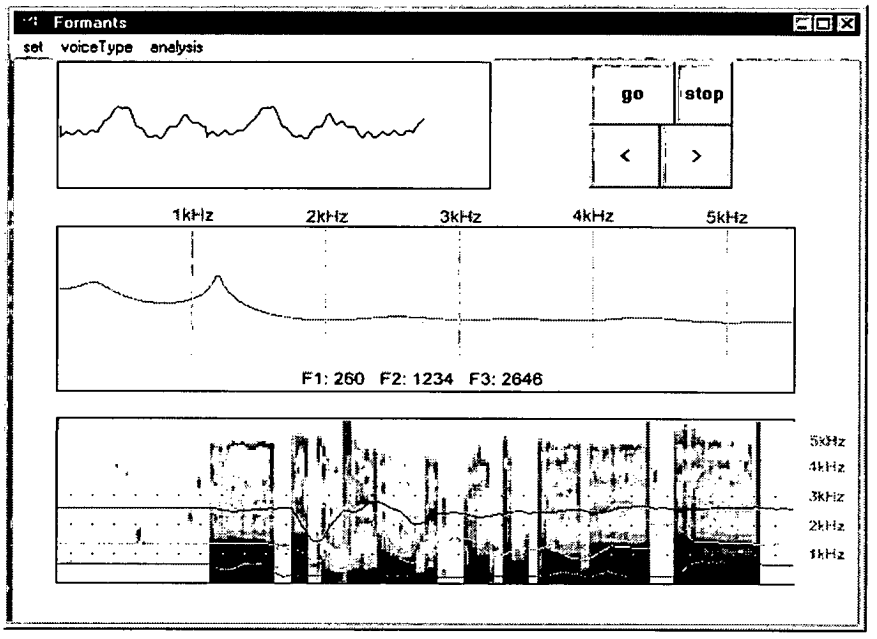


Figure 6.25. *Therapist's* graphical output of the real-time formant tracker. The top left window shows the current waveform; the middle window shows a slice of the LPC spactra; the lower window shows the grey-scale spectrogram, with the estimates of the first three formants overlapped.

Consonants

The problems related to consonant production discussed in Chapter 2 are generally encountered when trying to teach contrasts, in order to emphasise the difference in the articulators' positions. Therefore the following aspects of consonant production should be handled by the speech processing algorithms:

1. place of articulation contrast
2. voicing contrast
3. manner of articulation
4. initial and final consonants
5. consonant cluster

Place of articulation

Errors in place of articulation are common in deaf speech: that is, the movement of the articulators required to make the sound occur in the wrong place within the vocal tract. A common example hearing-impaired speech is the replacement of the palatoalveolar fricative /ʃ/ (as in "ship") with the alveolar fricative /s/ (as in "sea"), made further forward in the mouth. Software modules are required to detect and correct such errors, and also in some cases to help speakers to make smaller adjustments in the placement of their articulators.

Voicing contrasts

The contrast between voiced and voiceless segments is often reduced in the speech of hearing-impaired speakers, so that the distinction between /p/ and /b/, for example, is not made clearly or not made at all. The loss of this contrast can be extremely detrimental to the speaker's intelligibility.

Manner of articulation

Errors of manner of articulation include the replacement of fricatives with stops: this involves the use of a complete closure of the vocal tract, instead of the creation of a narrow air channel, and reflects the difficulties hearing-impaired speakers have in the co-ordination of speech movements.

Initial and final consonants

The omission of consonants completely is particularly common at the beginning and endings of words. This can cause intelligibility problems, since many final consonants in English, for example, are important markers of grammatical information such as tense and number.

Consonant clusters

The articulation of consonant clusters, such as /st/ or /sp/, may also cause problems. The omission of one of the consonants is a common feature. Another distortion which is commonly seen is the insertion of a vowel between the members of a cluster, often rendering the word that contains the cluster unrecognisable.

These aspects of consonant production have to be dealt with using two different approaches to the design of the speech analysis modules. The first aspect, *place of articulation*, is often corrected in speech therapy by asking the user to produce a long /s/ and to try to change it while monitoring the quality of the consonant in real-time with an appropriate display. A method for automatic analysis of voiceless fricatives is based on the determination of the frequency and standard deviation of the centroid of the spectral distribution (Wrench et al, 1995). The two values can be displayed on a two dimensional display, permitting the separation of several perceptually distinct voiceless fricative realisations. This method was implemented in a simplified way, sufficiently accurate for the study of the /s/ - /ʃ/ contrast, giving a uni-dimensional-only result of the analysis (see Figure 6.26). The speech input was sampled at 22 kHz and a 256 point FFT was calculated on a 97% pre-emphasised signal. The centroid of the power spectra was calculated approximately every 0.1 seconds, and the result was sent to the graphic routine for display. Components below 500 Hz were excluded by the calculation, in order to reduce the effect of breath noise. Furthermore, speech signals having a centroid of spectral distribution below 2 kHz were considered outside the range covered by the fricatives of interest. A typical output from this module is a gradual change from a value of about 4 kHz for a /ʃ/ to 7.5 kHz for a /s/.

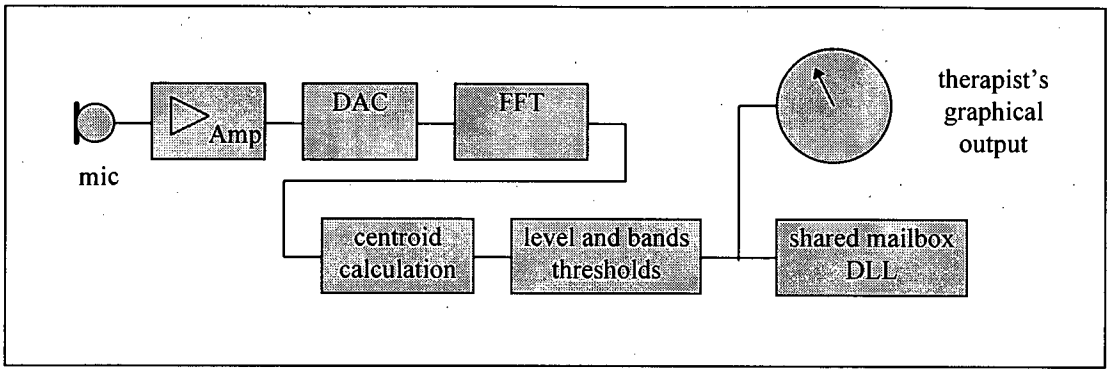


Figure 6.26. Voiceless fricatives analysis

Figure 6.27 shows the *technical graphic output* of the voiceless fricative analysis module. The spectral slice of the input waveform is overlapped to a marker showing its centre of gravity.

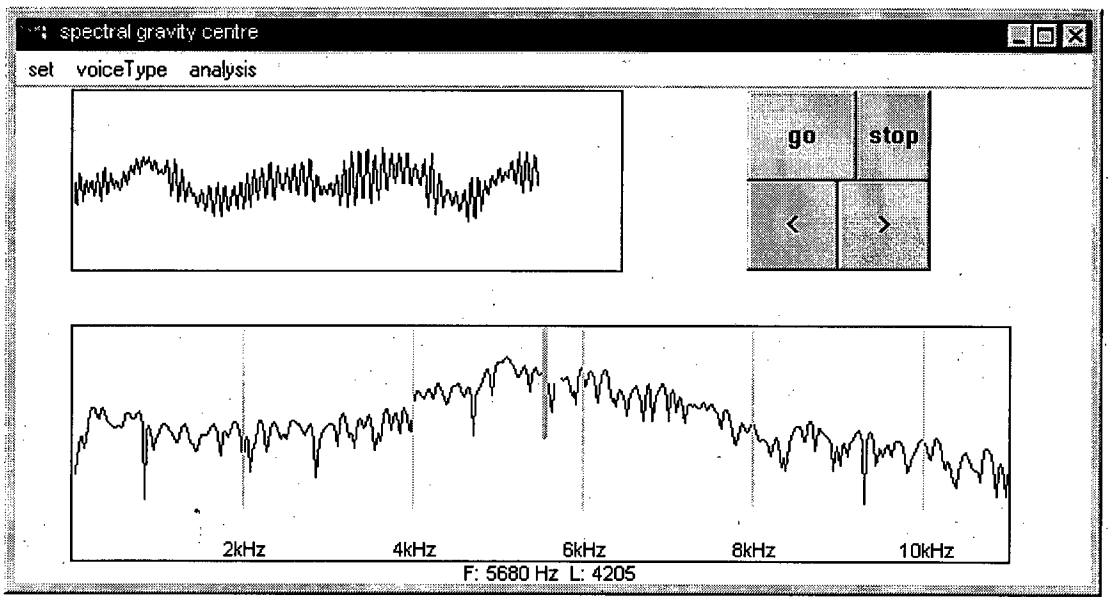


Figure 6.27. *Therapist's* graphical output of the voiceless fricatives analysis. The upper windows shows the current waveform; the main window shows a FFT spectra slice with a marker on the centroid of the spectra (5680 Hz in the example).

A different approach had be used for all the other aspects of consonant production, (voicing contrast, manner of articulation, initial and final consonants, consonant cluster) since these do not involve production of exclusively continuous sounds. A segmenter is necessary, allowing the system (and the user) to focus on specific sounds, or types of sound, even if they are embedded in words or complete sentences. The segmentation algorithm should be capable of at least broad class segmentation, and of running in near real-time. The block diagram in Figure 6.28 highlights the main procedures required for automatic phonetic segmentation.

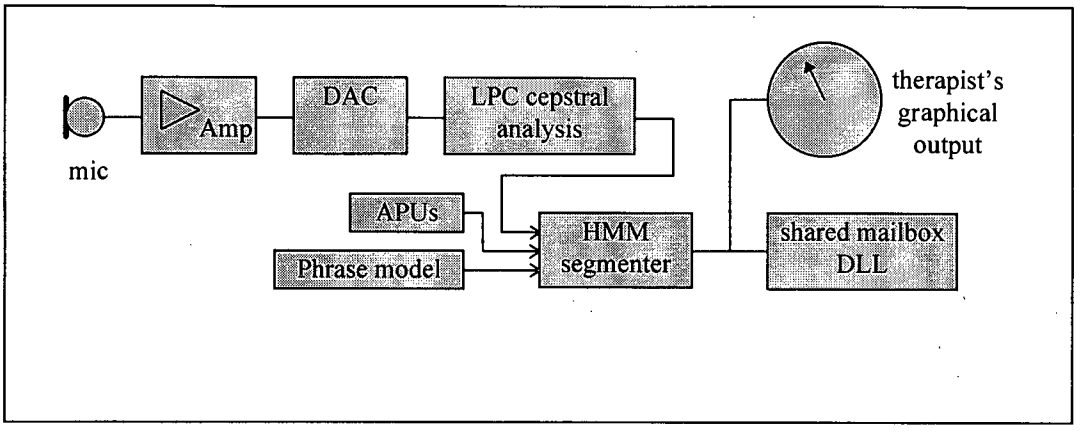


Figure 6.28. Automatic phonetic segmentation

A digitised speech waveform is end-point detected and input to the LPC cepstral analysis module (Markel & Gray, 1986). This module processes the speech in frames of 20 ms. with a shift rate of 5 ms. Each output frame consists of a vector of the first 10 cepstral coefficients including the zeroth coefficient which has been weighted by a scaling constant.

The segmenter itself has to work in near real time. Therefore computational efficiency was a major consideration in its design. Accordingly, a simplified form of hidden Markov model (HMM) representation was adopted, in which each of a suitably defined set of acoustic-phonetic units (APUs) is represented by a one-state model with a single cepstral centroid vector.

The APUs are a mixture of phonemes and portions of phonemes, together with one unit corresponding to silence. The rationale of the APU set design is that each APU should be representable, to a first approximation, by a steady state with a single cepstral target. To achieve this, phonemes with definite temporal structure (e.g. stops, affricates and diphthongs) are split into smaller units.

The APU models are trained on hand segmented speech. The acoustic representation used consists of the zero-th to 9th linear predictive cepstral coefficients computed in a 20 ms. window every 5 ms., with the zero-th coefficient multiplied by an empirically determined scaling constant (here set to 0.25). For each APU, the cepstral centroid is obtained by averaging together the weighted-mean vectors for all training segments of the appropriate APU identity, where the weighting within each segment is by a raised cosine function so that vectors near the centre of the segment are given most weight. The other parameters estimated for each model are a scale factor for the cepstral distances (which is made inversely proportional to an estimate of the mean squared Euclidean distance from the centroid); a self transition probability; and a gamma distribution for segment duration.

The possible pronunciations of the word or phrase to be segmented are represented by paths through a *phrase model* or pronunciation network. This is derived automatically from an orthographic transcription of the phrase together with phonemic representations of the words (which may include multiple pronunciations). Phonological effects such as reduction at word boundaries, and expected errors are incorporated. The network is minimised in the sense that nodes with the same predecessor or successor set and the same APU label are conflated.

The segmentation proceeds using a Viterbi algorithm (Rabiner et al., 1985), in which the best matching APU sequence from the phrase model is found together with its best-scoring alignment to the cepstral vector sequence representing the utterance. For each APU, the scaled squared Euclidean distance between the centroid and the observed vector for the current frame of the utterance is taken as the negative log emission probability (this corresponds to using a simple form of multivariate Gaussian distribution), and the negative log self-transition probability is taken directly from the model. When a transition between APUs is made, an adjustment is made to the accumulated negative log probability to convert the exponential duration probability distribution for the current APU, implicit in the basic HMM formulation, to the estimated gamma distribution. Such a post-processing form of duration modelling has been found to give similar results to an exact formulation, with much less computation (Rabiner et al., 1985). The implementation adopted imposes a maximum duration on each APU. Silence, however, is represented by an unconstrained duration model, in which the segment duration is unlimited and the transition and duration log-probability terms are set to 0.

The segmenter's output is a series of segments, each with start and end times and an APU label.

Implementation

The segmenter, (originally developed for intonation teaching purposes) was initially implemented sharing the computational load between the numerical accelerator board (performing the LPC cepstral analysis) and the PC's CPU, performing the segmenter task itself. After optimising the code, and with the availability of faster CPUs, the full algorithm code was moved to the PC's CPU. The performance of the segmenter and its speed greatly depend on the complexity and quality of the model used (McInnes, Carraro, et al., 1992). For a typical case where a few words such as "bell-spell" or "tea-sea" were used to study the contrast in the first consonant, the total processing time (after the utterance was end-point detected) was about 1.2 seconds (on a Pentium 90 MHz).

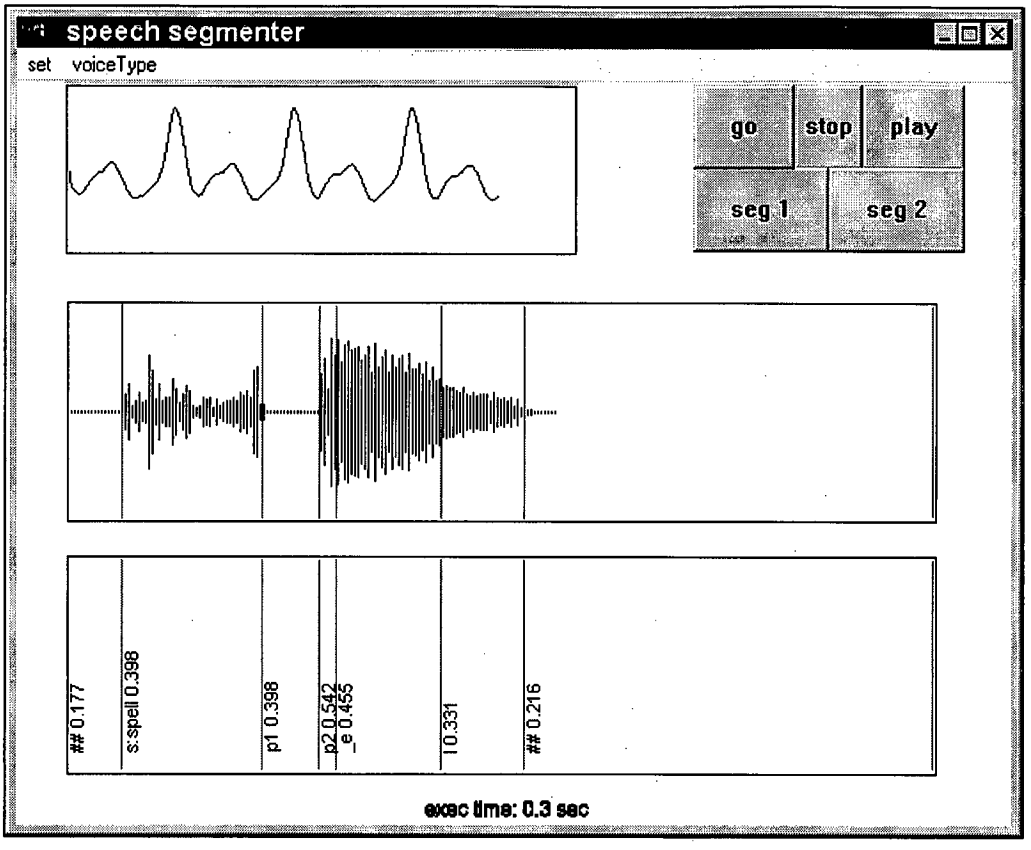


Figure 6.29. *Therapist's* graphical output of the speech segmenter. The top window shows the curent waveform; the middle window shows the amplitude envelope of the utterance; the lower window shows the segmentation results, aligned with the envelope display, in the form of the Acoustic Phonetic Units (APU) recognised, and the relative error.

The segmenter was written entirely in C using Microsoft Visual C++, and as with pitch tracker and the formant tracker, was realised as a independent programme having a "therapist's" graphic output consisting of three windows. The first one shows in real-time the input waveform, the second one shows a end-point detected envelope of the waveform, and the third one shows the segmentation labels (see Figure 6.29).

6.3.3 Graphics

The graphics were implemented using different tools:

- Microsoft C++ version 1.52 with the Graphical Device Interface (GDI) library
- Microsoft Visual Basic version 3.0 and 4.0
- Borland Delphi version 1.0

The three tools are different in terms of ease of use, flexibility and performance.

The GDI library called by C/C++ code requires the largest programming effort, since any graphical element has to be defined by lines of source code using low-level graphic primitives. Adding a window or a button means writing many lines of code, and the result is visible only after compilation. However there are no limitations in the graphic possibilities (even if some complex graphic function may require a very large programming effort), and an application written using this tool gives the best performance in speed.

Visual Basic, on the other side, is very easy to program, since the graphic elements are chosen from a toolbox, and they can be immediately positioned on the screen. Some of the properties of the elements, such as size and colour, can be easily changed without the need to write any code (see Appendix B). Visual Basic automatically generates empty subroutines for all the possible properties of the new element. These subroutines can be filled with code in order to tailor the behaviour of the graphic elements. However the simplicity of use has to be paid in terms of reduced speed (the language is interpreted) and some limitation in flexibility.

Delphi provides almost the same ease of use as Visual Basic, but since the program is compiled, it runs faster than the Visual Basic interpreted code. The drawback is that the size of the code is generally much bigger than a Visual Basic application.

The three tools were used for different requirements. All the "technical" displays were written using C and calls to the GDI library, since the large number of pixel-related operations would make impractical the use of both of the other choices. For example, the continuous update of the waveform in the top left window of the display shown in Figure 6.25 was impossible to achieve using Visual Basic. Similarly, the grey-scale spectrogram shown in the same picture requires several operations for every pixel displayed. The use of an interpreted language would have slowed down the display by a factor of 30 to 100 times. Delphi is much faster than Visual Basic in this type of operation, but since the technical displays were implemented as part of the speech processing modules, it was chosen to write them entirely in C, to assure the fastest execution time possible.

Visual Basic and Delphi were appropriate when animation was involved, since several animation functions are available in their libraries. These functions run very fast since they are not interpreted, but instead executed on optimised external modules. The animation functions include movement of *sprites* (arbitrary shaped areas) on the screen, optionally with collision control, overlapping of different planes, and resizing of objects. All these functions were exploited in the graphic feedback. Other multimedia functions, such as the playing of video clips in Audio Video Interleaved (AVI) format were achieved using the Multimedia Control Interface (MCI) of the Windows operating system. A special case was the Video Help function, were the more convenient MPEG video format was used[1]. This format allowed the playback of video clips with an adequate resolution and frame rate without the need of excessive disk storage. This video format is now standard in the last release of Windows (Win95 and NT-4). At the time of implementing the software, a third party software player was used, extending the MCI standard driver set.

6.4 Conclusion

A set of visual feedback technique for hearing-impaired speech rehabilitation has been designed and implemented, based on the results of experiments and on therapists' recommendations. This set of visual feedback modes tries to improve the effectiveness of traditional techniques, and proposes new ones never tried before. Furthermore, optimisation of real-time code, and low cost issues resulted in speech analysis modules that do not need specialised hardware to run on, allowing the use of this software on any recent higher specification Personal Computer or laptop.

[1] MPEG (pronounced M-peg), which stands for Moving Picture Experts Group, is the name given to a family of International Standards used for coding audio-visual information in a digital compressed format. The MPEG family of standards includes MPEG-1, MPEG-2 and upcoming MPEG-4, formally known as ISO/IEC-11172, ISO/IEC-13818 and ISO/IEC-14496.

CHAPTER 7

User Trials

7.1 Introduction

Evaluation of the prototype system developed in this research was an important phase where the work done was assessed with real users. The lack of adequate evaluation has always been a problem in the development of speech training aids. Therefore, in order to ensure that the prototype system reflected the needs of users, an assessment of the suitability and efficacy of the system was carried out with the help of hearing-impaired speakers and therapists, at several locations in Britain. The first phase of the evaluation overlapped in time with the design and implementation of the system itself, and this interaction resulted in many refinements to the visual interface. The activity consisted of four sets of evaluations with the following types of hearing-impaired subjects:

- Pre-lingually disabled
- Post-lingually disabled
- Elderly disabled
- Cochlear Implants

Pre-lingually disabled

This activity dealt with trials with groups of pre-lingually disabled speakers: that is, speakers whose hearing loss occurred in early infancy (or before birth), before the development of language. The evaluation ran for approximately 3 months. Both children and adults fall into this category, but therapists suggested that preferably older children would be suitable for therapy (and participation in

the trials), since younger children often have major language impairments and receive a different form of therapy (concentrating on language development and voice use rather than articulation and intelligibility). This group typically shows the greatest degree of speech impairment, and presents some of the greatest challenges to the feedback systems.

Post-lingually disabled

This activity involved groups of post-lingually disabled speakers: that is, speakers whose hearing loss started after the acquisition of language in infancy. Their speech impairment varies significantly according to the age at which hearing loss begun. There are not many people in this group, since there are now relatively few post-lingually deafened adults who have not already received cochlear implants. The evaluation ran for 2 months.

Elderly disabled

This activity involved a cohort of elderly hearing-impaired speakers, whose speech problems typically cover the areas of pitch, amplitude and duration rather than the segmental problems associated with other groups. The evaluation ran for 4 months.

Cochlear Implants

This activity involved trials with groups of speakers with cochlear implants, both children and adults. These speakers have some form of auditory feedback, but they have to learn how to interpret this feedback and use it to monitor and control their own speech. The evaluation ran for 4 months.

7.2 Aims and Structure of the trials

The work carried out on the evaluation of the prototype system was divided into 2 stages:

- a period of evaluation of the system while it was being developed, in close interaction with the users by means of consultations with them.
- a set of exploratory performance trials of the system.

This two stage approach to the evaluation of the system was useful since it allowed a certain amount of flexibility during the development of the system, while at all times maintaining a close focus on users' requirements and needs. This informal field trial did not involve comparison with other known systems or the use of a control group having standard therapy sessions.

7.3 Procedures and Results

Consultations with Users

As explained in Chapter 6, during the design and implementation of the prototype system, consultations were held with speech therapists, hearing therapists, teachers of the deaf and hearing-impaired users themselves to discuss the results of the experiments and how these results guided the development of the prototype system. These consultations concentrated on issues such as the nature of the feedback to be offered, and the degree of autonomy the system might be capable of providing. These consultations yielded a number of important recommendations for the design of the system. In the critical area of autonomy, for example, it was concluded that the system should function principally as an aid for therapists, but with provision for additional unsupervised practice on items which have been covered inside their normal therapy sessions. A major issue was the user's desire for much greater flexibility than is provided by existing systems, for example in the design of the user interface, the nature of the phonetic targets, and the choice of examples and analysis settings for different speakers. The discussions also touched on the question of the integration of a system like this into speech therapy provision. In Britain in particular, the nature of speech therapy provision is changing, with more and more therapists now required to go out to their clients (e.g. in schools) rather than working from a permanent, central facility. This confirmed that a portable system could be particularly valuable. Many of the suggestions and recommendations received at this stage were progressively implemented in the design of the user interface and the speech processing modules. Other enhancements to the system which were suggested - but which have not yet been implemented - included the provision of auditory feedback for the pitch, amplitude and vowel modules; the facility to have two microphones, one for the therapist and one for the client, to avoid the disruption of having to transfer the microphone when the therapist needed to demonstrate a sound; and the ability to store the speaker's best effort on the system to act as a target for future attempts.

Evaluations

The prototype system was evaluated in clinics and schools by therapists and their hearing-impaired clients. These evaluations were exploratory in nature, in order to gather information about the system and to implement improvements in response to users' requests. They concentrated on aspects of system use such as the degree of training required to use the system, its suitability for both supervised and unsupervised teaching, the acceptability of the user interface, and the usability of the system in general (e.g. robustness, clarity of instructions, handling of errors).

The first result of these evaluations was encouraging positive feedback on the potential capability of the system. Users felt that it was very motivating and exciting, and excellent for group work as well as with individual clients. The second result was a number of suggestions for improvement and

expansion of the system, covering issues such as navigation around the system (for example, a difficulty in setting the pitch scale for individual clients resulted in the simplification of the setting modality in all the pitch modules, and was also applied in the loudness modules); the suitability of the system for younger children; alternatives to the headset microphone chosen for the prototype; new ideas for the teaching of rhythm; the possibility of using the system with a group of children; and the importance of home use in maintaining the speech skills of older children once they leave school and lose contact with the speech therapy services.

In response to these suggestions, many improvements to the user interface were implemented. Additions to the system included the expansion of the vowel module to cover different accents, the provision of alternative forms of vowel feedback, increased control of the user interface, and most importantly from the point of view of flexibility, the introduction of basic authoring facilities for therapists in the pitch and vowel modules. The evaluations also showed a significant demand for visual feedback systems like this in the treatment of other speech disorders, such as dysarthria, aphasia and phonological impairments. Some therapists had already started using some of the modules for the speech rehabilitation of a dysarthric speaker (without hearing impairment), with some promising results.

Performance Trials

The final stage was a set of user trials in which the system was tested at a number of sites in Britain. Therapists used the system with a large number of clients, for some weeks, to complement and extend their normal therapy sessions. They provided feedback on both the efficacy of the system and any practical difficulties with its use. The main aims of these trials were:

- to gather data on the user's reaction and usability of the system in daily use in speech therapy clinics and schools;
- to investigate to what extent the system met the requirements of therapists and hearing-impaired users in the four user groups.

In contrast to the preceding exploratory phase, at this stage no changes to the system were undertaken. The trials concentrated on a range of issues, which can be grouped into four areas of interest:

- whether use of the system led to any improvement in the client's speech performance and intelligibility, as judged by the therapist;
- the effect of the system on users' motivation and self-confidence (identified as two of the chief objectives of speech therapy during early consultations with therapists);
- the client's sense of enjoyment (or conversely of boredom and frustration) at using the system;
- and the general usability of the system as a whole.

It was decided to base the trials on the subjective assessments of the therapists about the progress and achievement of users, rather than attempting to take objective measures (such as vowel quality or fundamental frequency measurements). This was due to pressures of time and the conflicting needs of the therapist. Assessment of the clients' performance in these areas, and of the system's general usability, was therefore performed using a set of questionnaires designed specially for the purpose. The aim was to allow each speaker to use the system for at least 4 sessions (typically once a week), in order to give them time to get used to working with the system, and to monitor any improvements or deterioration in their performance over this period. However, in most cases such extended use was not possible, and some of the users attended only a single evaluation session.

7.3.1 Trial locations and speaker groups

The user trials took place in the following sites:

- **Department of Speech and Language Therapy, Paisley Health Centre, Paisley**

  Therapist: Fiona Morrison

  A single speaker (a boy of seven years), profoundly pre-lingually deaf, evaluated the system for three sessions.

- **Nottingham Paediatric Cochlear Implant Programme, The Queen's Medical Centre, Nottingham**

  Therapists: Jayne Inscoe, Sarah Allen

  This Nottingham Centre specialises in the provision of cochlear implants to very young children from all over Britain, and provides an extended therapy service after surgery. The NPCIP evaluated the system with a total of nine such children, all pre-lingually deafened. Ages ranged from 2 to 8 years. With these subjects, it was possible to have only a single evaluation session in each case, since the children being treated under this programme attend the central clinic at the NPCIP only once or twice a year, receiving their regular therapy at their local clinic.

- **Dept of Speech and Language Therapy, Dundee Royal Infirmary**

  Therapists: Susan Howden and Fiona McHugh

associated with the elderly), while the third had a cochlear implant. It was therefore decided to include these speakers in their corresponding groups.

7.3.2 Procedures

For each trial, the system was set up in the clinic or school that participated to the evaluation, and left there for the duration of the trial. A detailed demonstration of the operation of the system was given to the therapists involved, and they were trained in its basic functions. They were provided with session log sheets and evaluation questionnaires, and with instructions on which parts of the evaluation to perform at which time.

The selection of the clients was left to the therapist's judgement, as was their choice of which modules would be appropriate for evaluation. The idea was that the therapists would be able as much as possible to integrate the use of the system into their normal therapy session. This policy was adopted to avoid any possible ethical problems which might arise if therapy sessions were restructured to suit the conduct of the trials: although in all cases the clients were taking part in the trials willingly, it was felt by the therapists themselves that the clients' needs and their right to effective therapy during the limited time available should always have the priority.

Therapists used the session log sheets during and after each therapy session to record the modules that had been attempted, any problems which occurred with the system, and any observations on their progress during the session.

A set of evaluation questionnaires (reported in Appendix A) was used by the therapists and their clients at various points in the user trials:

- During and/or immediately after each session, the therapists were asked to complete an evaluation of the client's progress with the modules that were used, to assess the client's confidence and motivation, and the performance of their speech during the use.
- Once each speaker had completed his or her set of sessions, they were asked to complete (with the therapist's help in the case of young children) their own evaluation of the system.

The questionnaires use a 5-point Likert scale (*Strongly disagree - disagree - neutral - agree - strongly agree*) to assess a number of aspects of the use of the system in a way that can allow comparisons between user groups and the individual modules of the system. Therapists were also encouraged to write comments and observations, and many chose to present these in the form of an small report.

The system was trialled at three sites during this time: a primary school unit, a secondary school unit and a health clinic for adults (post-lingually deaf). The system was trialled with twenty one pre-lingually deaf children, ranging in age from 3 years 9 months to 16 years old; however, repeated sessions were possible with only five of these speakers, giving a total of 31 sessions in all. The system was also trialled with two post-lingually deafened adults (aged 49 and 81 years) at the health clinic.

- **Yorkhill NHS Trust, Glasgow**

  Therapists: Kim Davidson-Kelly and Julie McCracken

  The system was evaluated at two sites (Possil Park Health Centre and Park House School for the Partially Hearing). Five children, ranging in age from 5 to 14 years, used the system over a period of nearly three weeks, giving a total of eleven sessions in all.

- **Royal Southampton Hospital, Southampton**

  Therapists: Sarah Worsfold (an adviser on hearing impairment to the Royal College of Speech and Language Therapists) and Sarah Paganga, Southampton University.

  Eleven speakers (two children aged 6 and 12; seven adults with ages ranging from 30 to 49; and two elderly speakers aged 70 and 76) gave a total of 32 sessions .

  The breakdown speakers into the 4 cohorts is shown below in Table 7.1.

Cohort	Adult male	Adult female	Child male	Child female	Totals
Pre-lingual	1	3	18	10	32
Post-lingual	1	6			7
Elderly					0
Cochlear Implant	1		4	5	10
Totals	3	9	22	15	49

Table 7.1. Breakdown of UK clients by hearing disability, age and sex

By far the majority of subjects were pre-lingually deaf children. There were small numbers of post-lingually adults, but no post-lingually deaf children at all in the sample. This reflects the fact that many children now receive cochlear implants, and therefore fall into the fourth category. There were only three elderly speakers present, but two had disorders which were more properly classified as pre-lingual deafness and post-lingual deafness (rather than the problems with presbyacusis normally

7.3.3 Results: general comments

The therapists provided a number of general comments and observations about the system. These covered points such as the suitability of the system for users of different ages and levels of disability, problems with understanding what the system was for, difficulties with the microphone or mouse, and any evidence of improved motivation or awareness of speech as a result of using the system.

With younger children e.g. D3 (a child aged 4 years) the system was used mainly to encourage vocalisation, since this is one of the main aims of therapy at this stage. It was felt to be very effective in this, even though this was not what the modules being tested (pitch and loudness) were designed for. On the other hand, some of the very youngest children (e.g. N1 aged 2) enjoyed watching as the therapists made different noises into the microphone, but would not try speaking into the system themselves.

Older children appeared to encounter fewer problems with the performance of the system, but one 16-year old (D17) felt that the graphics were more suitable for young children. One surprising point was that the oldest speakers had no problems with the system; speaker S11, for example, a 76-year-old woman, found the system very motivating and was keen to keep trying.

The possibility of home use was again raised, particularly in the case of speaker S2, who was very keen to demonstrate the system to his father and brother. It was also pointed out that the weekly sessions which were possible with the system were not ideal, and that daily usage would have improved confidence and independence.

Some of the therapists noted a positive carry-over into other therapy sessions: speaker G5, for example, showed "exceptional carry-over from previous session and to future sessions even once the system was removed [at the end of trial] - *very* exciting". In this case, the therapists started using pictures to remind him of what he had seen and done once the system had been removed. Another general effect was an improvement in self awareness and confidence: by his fourth session, for example, speaker S8 was freely commenting on his own performance and modifying it at the next attempt.

7.3.4 Results: individual modules

The therapists provided a great deal of feedback on the benefits and the drawbacks of individual teaching modules.

Pitch

The pitch modules were among the most used modules during the trials, and therefore a considerable amount of feedback, both positive and negative, was obtained. Pitch modules which were trialled included:

- the aeroplane game (with an option for split screen display)
- pitch speed videos
- pitch height/perspective videos

It was noted that displays in which pitch changes were represented by speed or by horizontal movements were less successful than those using vertical movements. The aeroplane pitch game was generally the most successful module for encouraging pitch changes: according to one therapist, speakers "quickly grasped the idea of raising and lowering the pitch", though one (equally quickly) learned that it was possible to avoid the obstacles by using a constant pitch at a level between the obstacles rather than varying the pitch to conform to the shape of the obstacles as was the intention of the activity. There were also difficulties with controlling the starting and finishing height of the aeroplane on the display[1]. Despite these difficulties, this module was one of the most popular and most adaptable, and its split screen facility was much appreciated. One therapist identified nine related activities for which this module might be suitable:

- voice on/off
- sustaining sounds (breath control)
- number of syllables produced on one breath (for speech rate and voice control)
- sustaining low/medium/high pitch
- controlling pitch movements (e.g. patterns representing typical word patterns)
- varying pitch movements continuously
- voice/voiceless contrasts
- copying therapist's model on split screen with sounds, words and phrases
- creating users' own "landscapes" to follow

[1] The authoring facilities included in the aeroplane pitch game as described in Chapter 6 were not completed at the time of the trials. The problems listed above are solved in the actual version.

The speed and height videos were seen as less successful, though one child (D10) was able to interpret the building height display without difficulty[1].

Intonation module

The intonation module was relatively undeveloped compared with the rest of the system, and this lack of maturity was apparent in the largely unfavourable reaction it received from therapists. Users were dissatisfied with the amount of feedback given, and with the time delay in its response. However, it was found to be very useful by one adult male speaker (S8), who was very pleased with the visual feedback it offered. This module clearly needs to be improved, and guidance must be offered to therapists as to when it is appropriate to use it and when not[2].

Loudness

In general the loudness modules were liked, and it was felt that they could give useful feedback and had a lot of potential. The following loudness modules were included in the trials:

- smiling face
- dog and bone
- ghost

The smiley face was well liked, in particular the idea of changing the graphics according to whether the voice was too quiet ("I can't hear you") or too loud ("Don't shout!"). The main problems with this module were its potential for giving negative feedback to some speakers, encouraging inappropriate loudness and shouting by the "reward" of a bigger face, despite the indications in the captions. One therapist pointed out that the module caused some problems for adults, who are told to watch the body language of listeners: this module gave inappropriate messages, since if a deaf person shouts the listener will move further away rather than closer as the face does. Several therapists pointed out that this module needs some means of reversing the display to cope with this and other problems.

The remaining modules were also liked, and were felt to be useful for different purposes, such as getting children to say "boo!" in the case of the ghost.

All the loudness displays were felt to move too quickly to register some changes in sound level, and did not work on running speech since they were designed only for sustained vowels.

---

[1] The *spinning ball* game described in Chapter 6 was not completed at the time of the trials, and was not included in the set of games.

[2] The intonation module (described as *delayed pitch feedback* in Chapter 6) was in its infancy at the time of the trials. The actual version works in near-real time, and the user interface has been completely redisigned.

Vowels

The vowel modules were also fairly extensively used in the trials, and received a very positive response. They were felt to be particularly appropriate for older users rather than for the very young children, as the display was still rather complex for them. Some therapists found the modules useful as a check, to confirm what they perceived subjectively.

In some cases the vowel modules were found highly motivating. Speaker G2, for example, became more and more intrigued as the session went on, and expanded the difference between some of the target vowels in response to the feedback. The moveable targets were found to be useful in reinforcing some of the client's less consistent contrasts. Some clients found it difficult to keep their attention on the task for more than a limited time. Therapists also requested some way of varying the vowel labels used; one used Cued Articulation as an aid, while another suggested a method of coding the vowels with key-words based on the names of colours which contain them (e.g. green for /i/, red for /ɛ/, black for /a/ and so on), a method which she has applied successfully with children.

It was felt (in the Scottish sites) that more Scottish vowel examples were needed, although therapists also saw the benefit of being able to move the targets to new positions to compensate for certain accents differences.

Some of the most encouraging comments came from one therapist in Dundee, who observed that she could "...see lots of potential with this programme - not just for deaf people". She felt that the module had been particularly valuable in one case (D14): "This boy likes the vowels, which are normally very hard work. This made it easier for both him and me!"

Consonant demonstrations

A range of animated sequences was included to demonstrate the formation of certain sounds such as /t/, /k/, /s/ and /ʃ/, without allowing speech input or any actual analysis. These were found to be interesting and to have great potential for helping clients to understand what was happening inside their mouth. Therapists commented that the vocal tract diagrams complemented well the information they themselves provided on consonant articulations during normal therapy sessions; they were particularly useful for place of articulation (less so for manner).

The /s/-/ʃ/ module

This module provided feedback on prolonged fricative articulation, distinguishing two fricatives which cause a lot of problems for the hearing impaired. Several speakers found this module helpful - one had not realised until trying it that she was pronouncing /s/ incorrectly, while another speaker made rapid improvements in the production of /ʃ/ in subsequent sessions. The analysis was not always sensitive enough to show all contrasts, but when it did work it was extremely useful. It was pointed out that the animation should show the rounding and protrusion of the lips which normally accompanies /ʃ/ production, as well as offering some reward as a means of motivating younger children.

The /s/-/t/ module

This module was thought to be a very good idea, and highly motivating for all children, with great potential for skill building and control of this and other contrasts. However, it caused a certain amount of frustration when it failed to work properly because of speech recognition errors. It was used only occasionally by the therapists because of these problems, and clearly needs some major improvements[1].

Help System

The video-based help system was also evaluated by some therapists during the trials, and was generally felt to be valuable, though limited in its usefulness by lack of content in some areas. The possibility of having sign-supported speech as well as subtitled speech was clearly appreciated by some clients. This element needs a fuller evaluation once the remaining details have been developed, to ensure that it performs its task properly.

[1] As discussed in Sections 6.2.4 and 6.3.2.3.4 the module *delayed feedback for consonant contrast embedded in words* was not giving results reliable enough because of the lack of good quality models for consonants. It was anyway in the set of modules used in the evaluation.

7.3.5 Evaluation questionnaires

The evaluation questionnaires which were completed by the therapists and their clients were designed to answer more general questions of usability and efficacy as indicated above.

Three questionnaires were administered in total (see Appendix A):

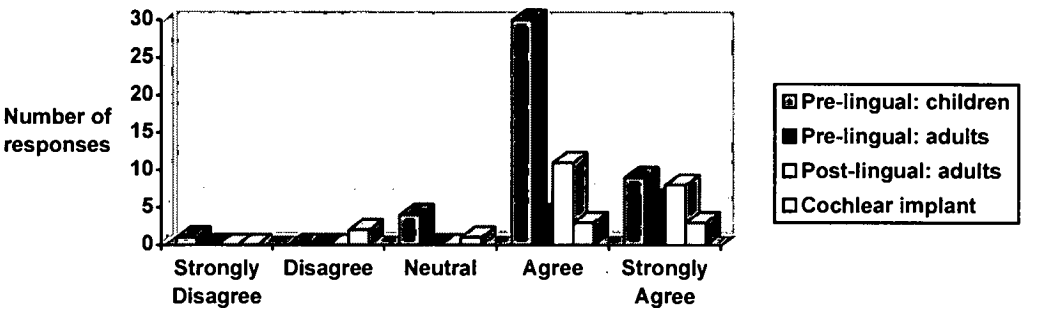
1. the therapist's evaluation of the client's performance with the system (a 10-item questionnaire)
2. the client's own assessment of the ease of use of the system (10 items)
3. children's own assessment of the ease of use of the system (5 items)

The results of each of these will be presented separately, analysing the responses to each question in turn. The responses are broken down into four groups, separating out the children and adults with a pre-lingual impairment, in order to see whether any issues particularly affected children rather than adult users. The results are presented here in graphical form, showing the actual number of responses (not normalised). No statistical analysis of any group differences has been carried out, given that most groups were fairly small and of unequal composition.

Therapist's evaluations of client's performance

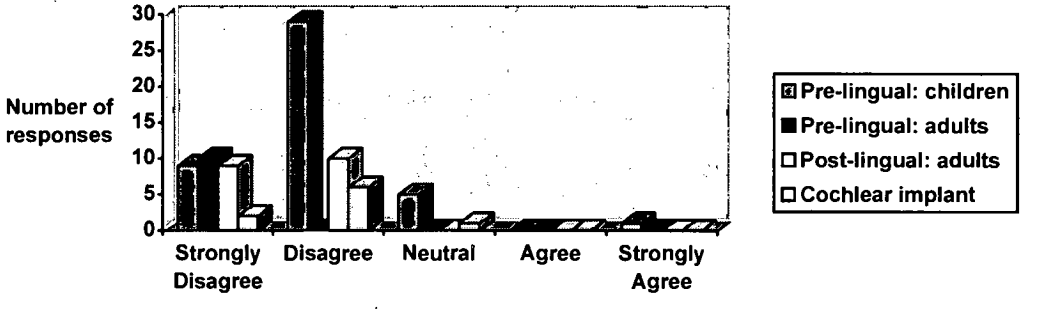
This questionnaire was intended to be completed by the therapist immediately after a session with a client. It contained 10 statements, covering both positive and negative aspects of the use of the system, with which the therapist could agree or disagree, providing their views on a 5-point scale.

Q1 The client found this system enjoyable to use:



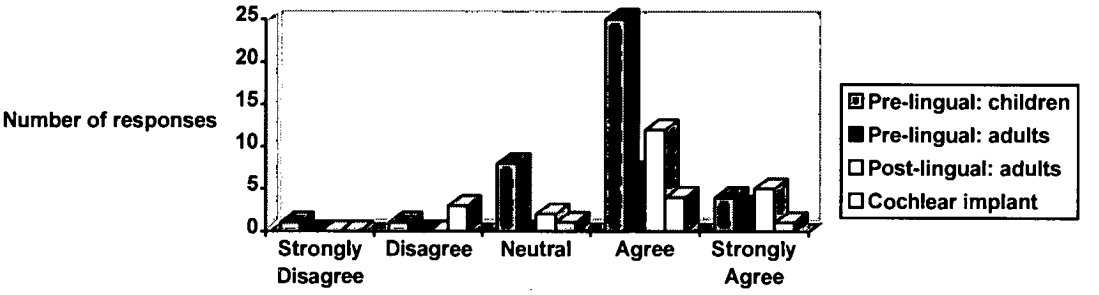
The responses to this statement were very encouraging, with most speakers in all groups expressing agreement or strong agreement with the statement. The cochlear implant users were more widely spread across the response categories, though the reasons for this are not clear.

Q2 The client found this system boring to use



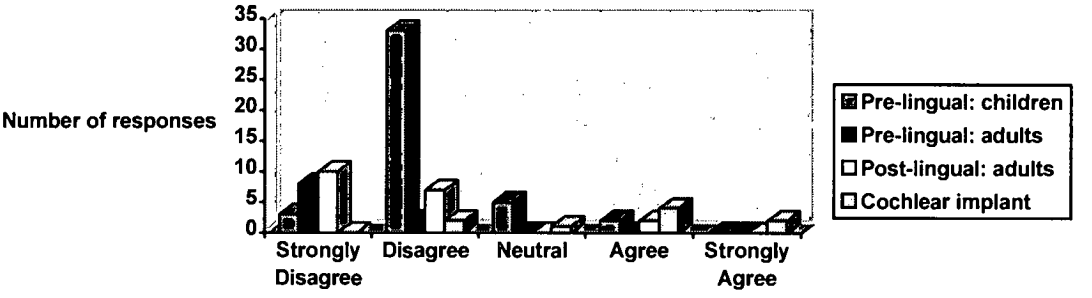
Again, there was largely a positive response to this point, with the majority of users disagreeing or strongly disagreeing with the statement.

Q3 The client was keen to use the system again



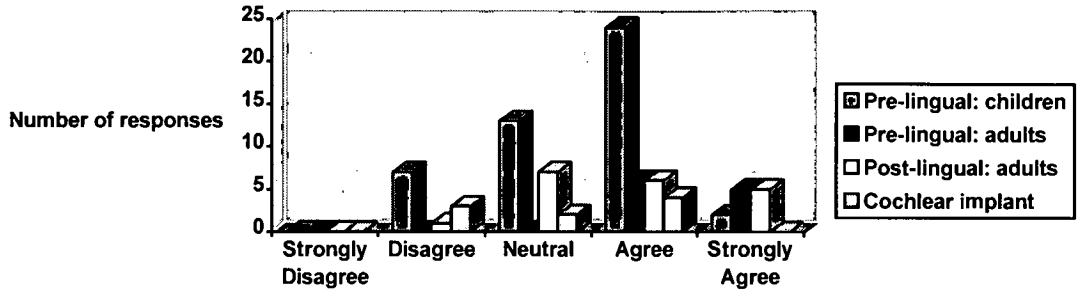
This positive statement also gave a very encouraging response, though this time with a number of pre-lingually deaf children having a neutral attitude.

Q4 The client found the system frustrating to use



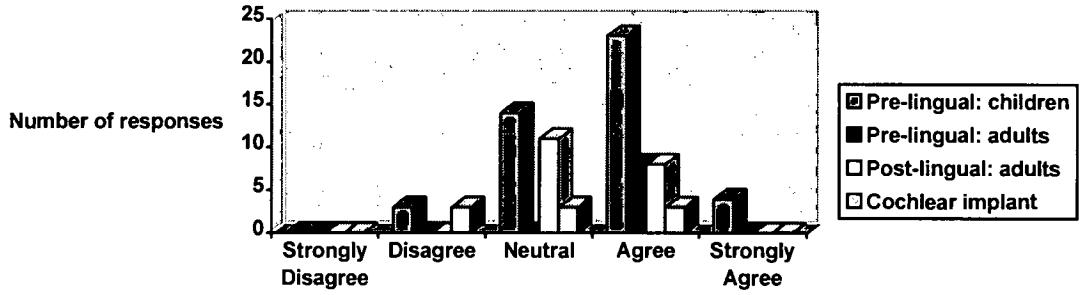
This statement brought a largely positive response, particularly for the pre-lingually impaired children and adults; however, there were significant numbers of cochlear implant speakers who experienced problems with frustration. This may reflect the fact that the system did not offer any auditory feedback, which they are capable of perceiving.

Q5 The system helped the client's progress towards the goals set for their therapy



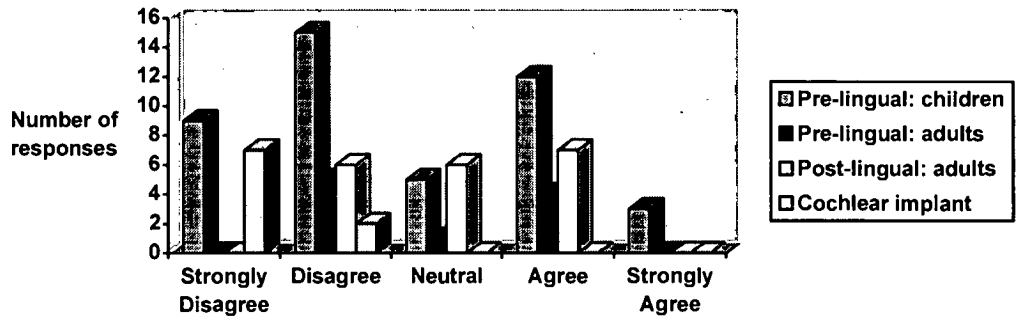
This statement received a more mixed response with several therapists dealing with the pre-lingually deaf children adopting a neutral or negative attitude. It is possible that the limited amount of time many of the clients (and therapists) had with the system - particularly for those clients who had only a single session - may have made this a difficult issue to judge. However, none of the nine therapists showed strong disagreement with the assertion, and 6 out of 7 with clients in the pre- and post-lingually impaired groups gave a positive response.

Q6 The system has made the client more keen to receive therapy



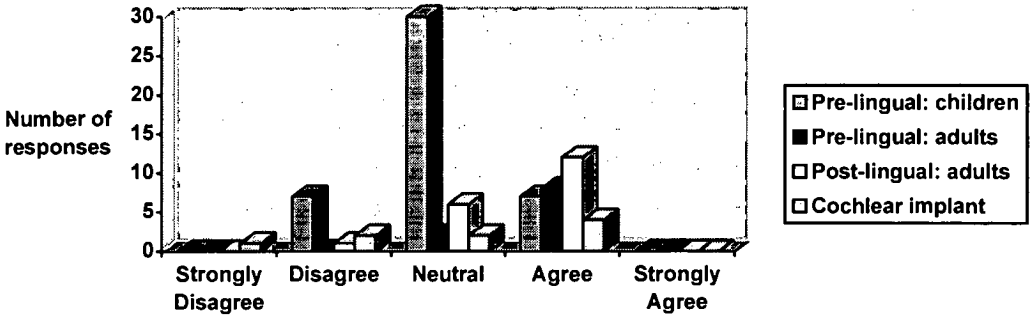
This statement too evoked a fairly mixed response across the four groups of users. Again the pre-lingually deaf children showed the greatest enthusiasm, with cochlear implant users being less influenced than the other groups. This ties in with their other responses, indicating that their experience of the system (in the opinion of their therapist) was less satisfying.

Q7 The client is ready to use the system by themselves



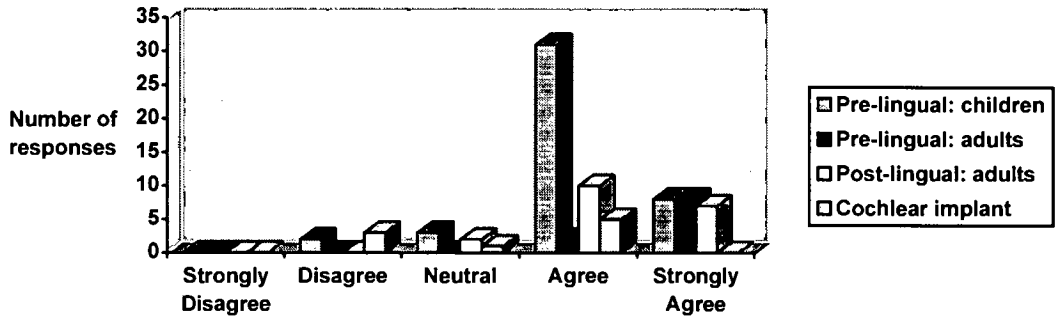
As expected, this statement drew fairly strong disagreement from therapists. Therapists are generally very wary of allowing clients to use systems such as this unsupervised, because of the dangers of negative reinforcement and demotivating feedback which are present when there is no-one there to interpret the feedback offered by the system. This said, there were still a significant number of positive responses which were very encouraging, suggesting that the system has reached a level at which at least some motivated speakers are capable of operating it autonomously.

Q8 The client's confidence in his/her speech production improved as a result of using the system



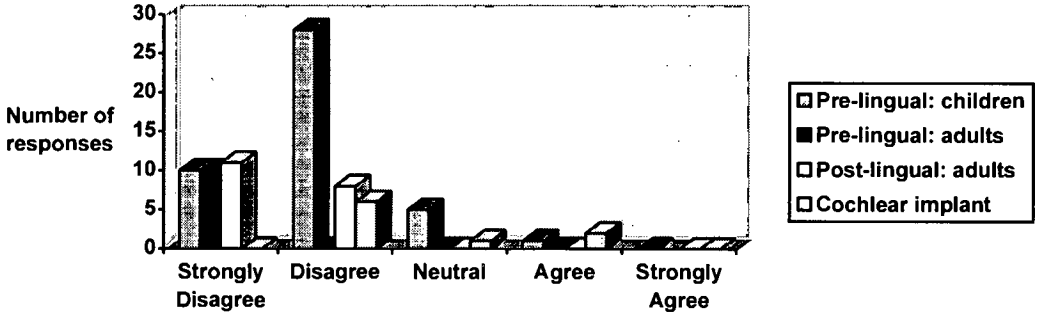
This question was included to assess the effects of the system on the client's confidence in their own speech production. The results suggest that there is a beneficial effect in some cases, but that therapists are neutral in many others, particularly in the case of children.

Q9 The client responded positively to the feedback provided by the system



The results from this statement show a strong positive response from clients to the feedback offered by the system, with therapists agreeing or strongly agreeing in the majority of cases. This trend is reversed for some children (both cochlear implant users and pre-lingually impaired), and a further examination of the data suggests that these speakers are among the youngest to try the system (some being as young as two years old). Further work on the visual feedback might usefully be undertaken to make it more accessible and motivating to these users.

Q10 The system made their speech performance decline

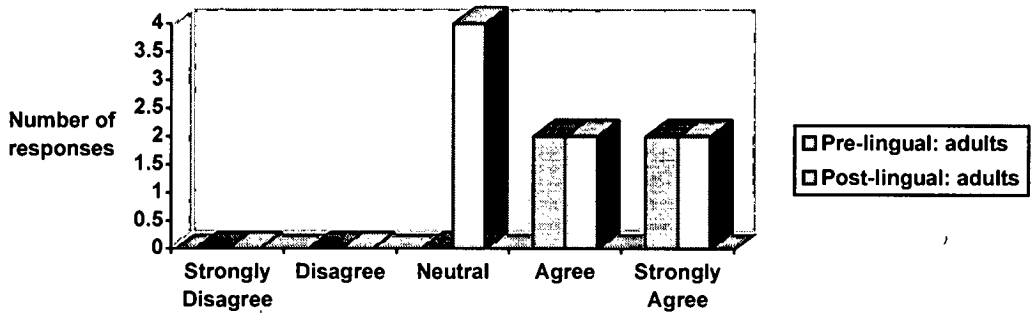


The final item in this questionnaire again gave encouraging results in these trials, with the vast majority of therapists disagreeing or strongly disagreeing, particularly in the case of adult speakers. A small number of cochlear implant users appear to have experienced some problems, however, possibly (as noted above) as a result of being deprived of usable auditory feedback or of being made to feel self-conscious.

Clients' evaluations of system usability

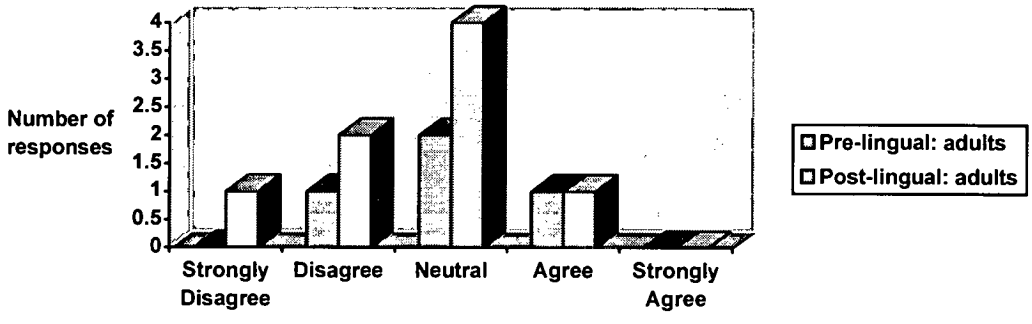
This questionnaire was intended to be given to the client after all their therapy sessions had been completed. It contained 10 statements, again covering both positive and negative aspects of the use of the system, with which they could agree or disagree using the same 5-point scale. This version of the questionnaire was designed for use by adults. A simplified version, Questionnaire C12, was prepared for use by children, and the results of this are presented later. Only two user groups (pre- and post-lingually impaired) are represented in this section of the evaluation, since the only other adult (a cochlear implant user) did not complete an evaluation.

Q1 I think that I would like to use this system a lot



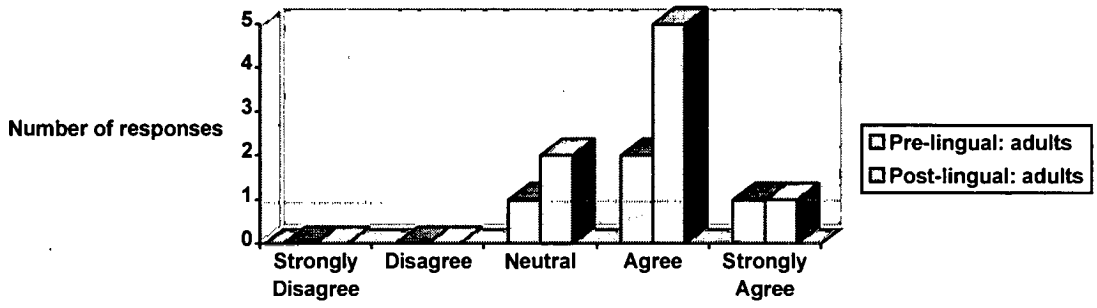
This first statement elicited a largely positive (or at worst neutral) response, with most speakers agreeing with it.

Q2 It took me too long to access parts of the system I needed to use



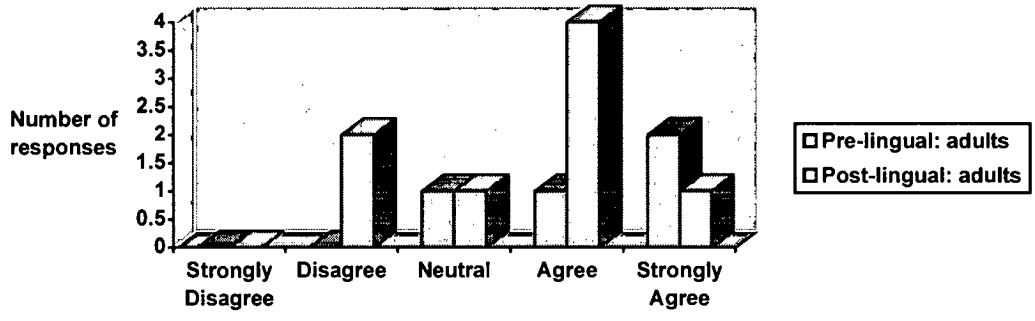
There was a more mixed response to this item, with at least one of the clients in each group agreeing; most either disagreed or were neutral.

Q3 I thought the system was easy to use



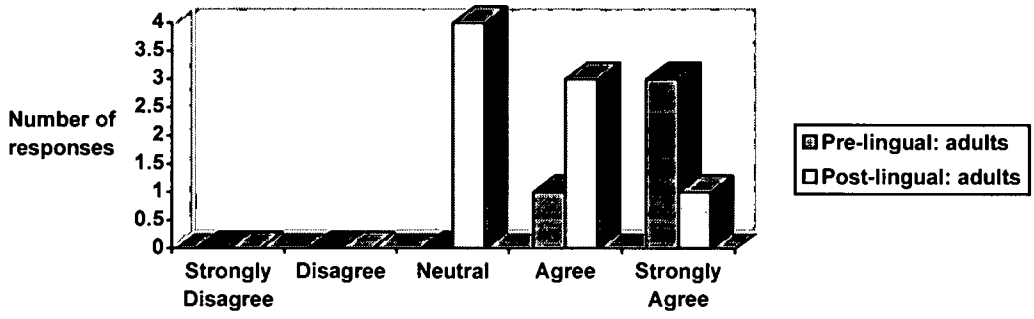
Most users agreed or strongly agreed with this statement. This was an encouraging result, suggesting that the care which had been taken to make the system user friendly had been well invested.

Q4 I think I would always need the help of a therapist to be able to use this system



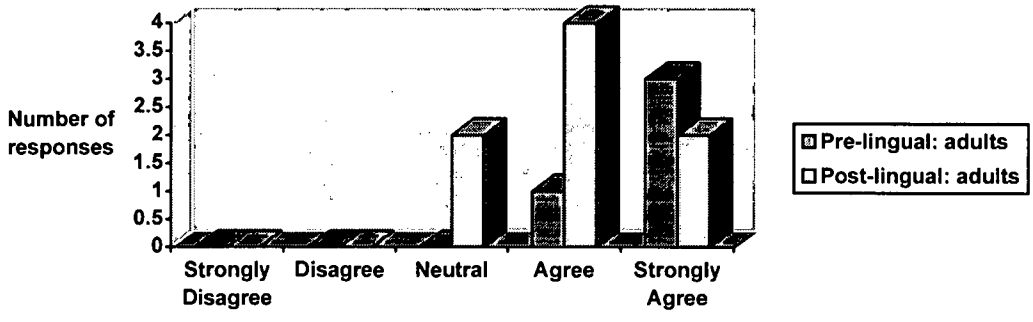
This statement was designed to test whether the perceptions of clients about the desirability of autonomous use of the system matched those of the therapists. From these results, it appears that most speakers do indeed prefer to have a therapist on hand.

Q5 This system gives excellent coverage of the speech features I need to work on



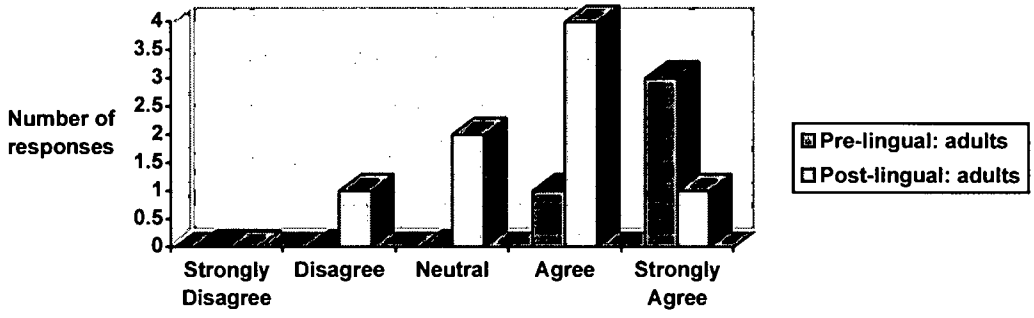
The positive response to this question, with most clients agreeing or strongly agreeing, suggests a reasonably high level of satisfaction with the courseware coverage.

Q6 I thought the pictures were easy to interpret



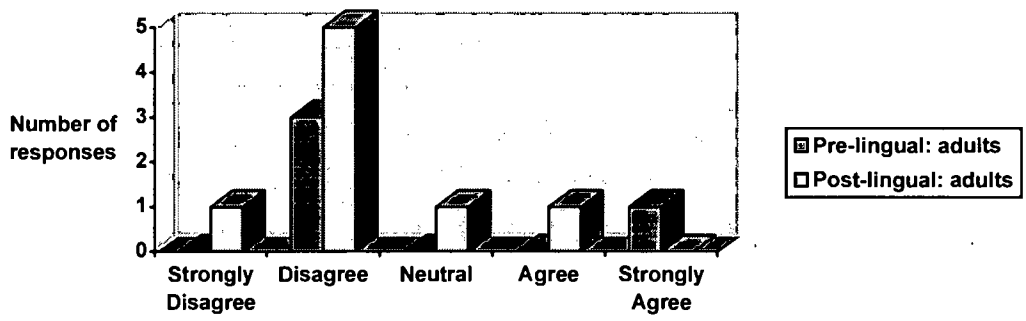
Again, this statement has generated a positive and encouraging response, with most speakers agreeing or strongly agreeing. This is despite the fact that some of the pictures presented to users are relatively complex. It appears that they do give adequate feedback for user to learn from.

Q7 Most people would learn to use this system very quickly



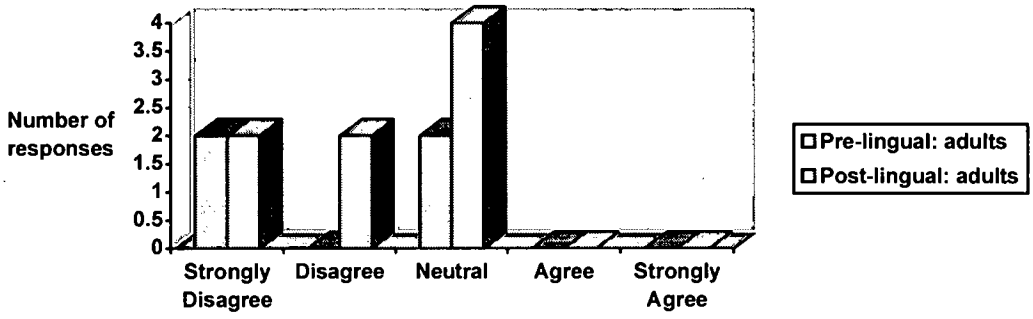
The response to this statement was also encouraging, suggesting that adult users see few problems in learning how to use the system.

Q8 I found it difficult to adapt the system to my needs



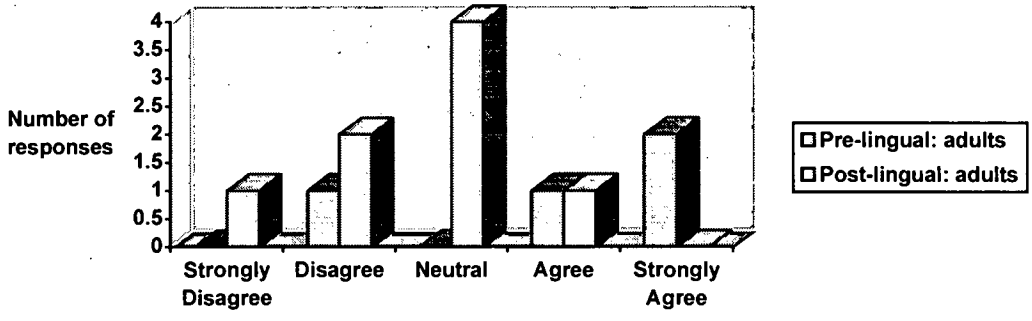
The issue of flexibility and adaptability was addressed in this statement; it appears from the data that most speakers had no such difficulty, though one or two did have problems.

Q9 I felt flustered using the system



This statement looked at the extent to which users could be intimidated or unsettled by the technology, but no problems were found, with both groups of users showing either disagreement with the statement or a neutral attitude. It appears that the level of control available to users is satisfactory.

Q10 I needed to learn a lot of things before I could get going with the system

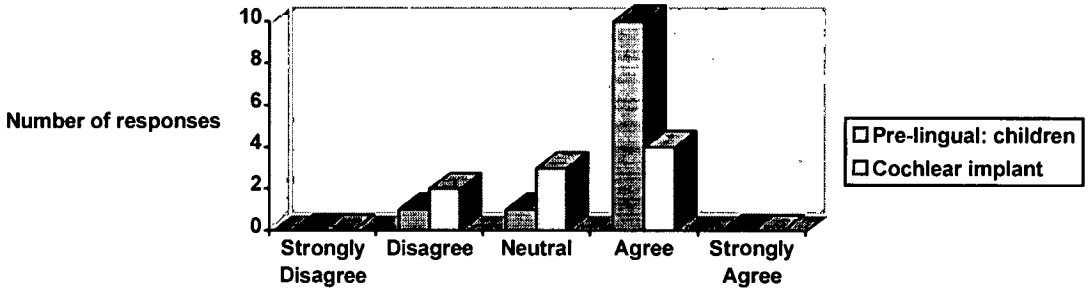


This statement attempted to measure the perceived complexity of the system, and the learning curve required to operate it. It evoked a mixed response, with many people showing a neutral attitude, while others did indeed think that a significant amount of learning and training was required.

Children's evaluations of system usability

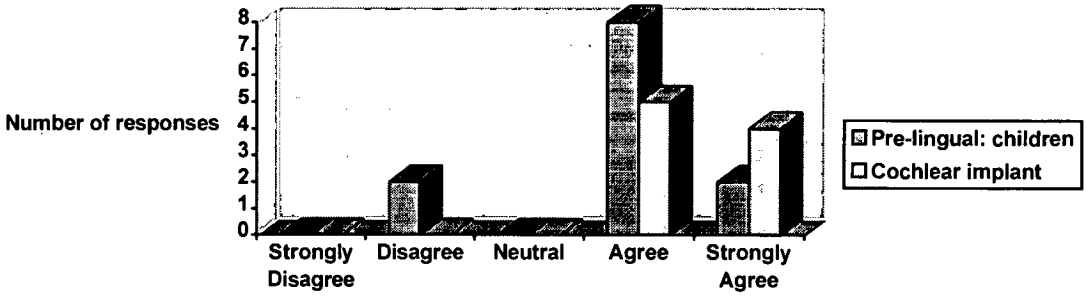
This questionnaire was intended to be given to the children after all their therapy sessions had been completed, to be completed with the help of the therapist if necessary. It contained five statements corresponding approximately to Questions 1, 4, 6, 7 and 9 of the adult version of the questionnaire (Cl1), again covering both positive and negative aspects of the use of the system.

Q1 I think I would like to use this system a lot



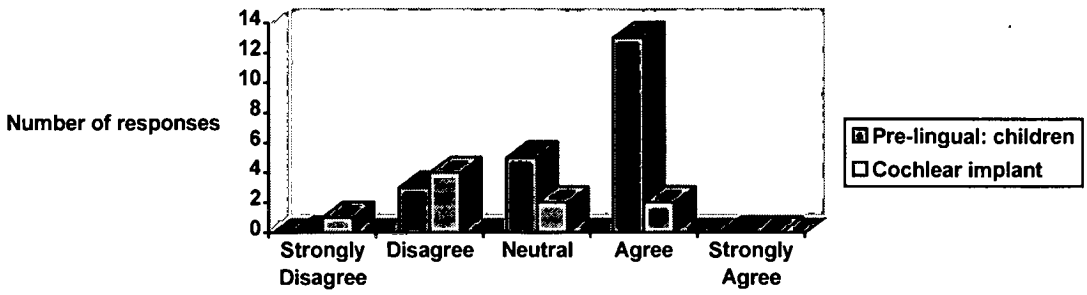
This question gave a mixed response, although one user was largely favourable to the system. Again, the cochlear implant users showed slightly more reserve in their judgement of the system, as has been seen in earlier analysis.

Q2 I think I would need my therapist to help me



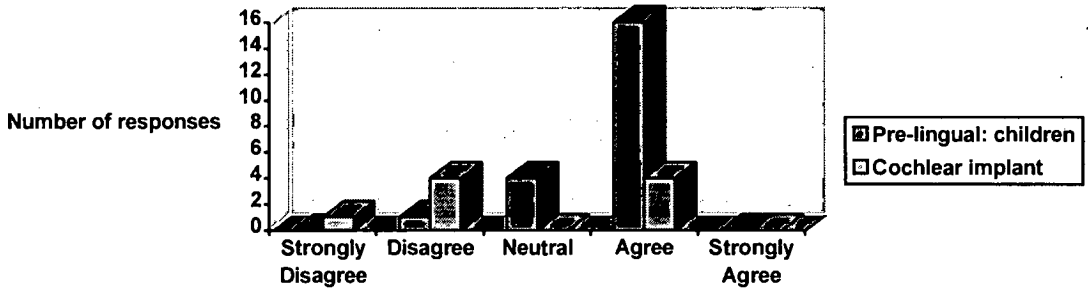
The client's response to this item clearly suggests that the children who have tried the system are not yet ready for unsupervised use, as their therapists indicated in Question 7 of Questionnaire Th2.

Q3 I knew what the pictures were telling me about my speech



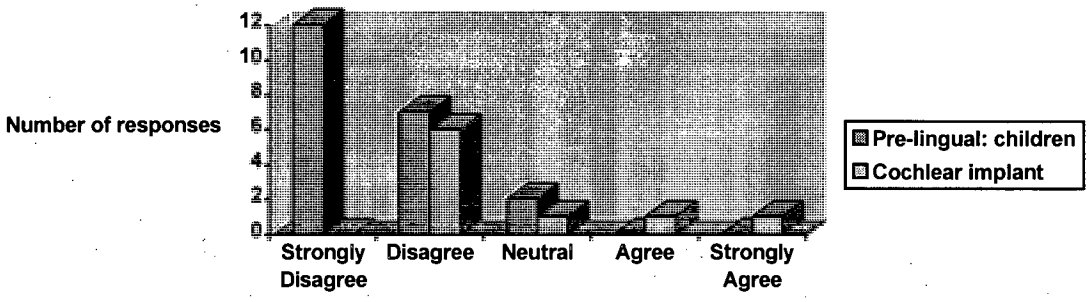
This rather mixed response also appears to support the therapists' indications that some of the children had difficulty in interpreting the output of the system. This suggests that some simplification of the graphics and the visual feedback would be required for successful operation with young children.

Q4 I think you could learn to understand the pictures quickly



This question, designed to see whether users found it easy to learn about the visual feedback, gave a similar response to the preceding item, with mixed opinions. Again, there are slight differences between the pre-lingually impaired children (generally more favourable) and the cochlear implant users.

Q5 The system made me feel nervous



In the light of the preceding answers, the high level of disagreement with this proposition was very encouraging. Users, even fairly young children, appear to be at ease with the system, making it more likely that they will eventually benefit from therapy.

7.4 Therapists' recommendations and suggestions for improvement

The therapists made a number of general and specific recommendations for improving the performance of the system, some of which are presented here.

One of the major recommendations was for a portable version of the system to be produced. A version which ran on a laptop PC, for example, would enable peripatetic speech therapists to offer the same extended service as those operating from central clinics, and would make the service available to those users (particularly elderly speakers) who cannot or will not travel for therapy and are seen in their own homes.

Many therapists and users found the mouse difficult to use, and therapists suggested either making the boxes and buttons on the screen larger and easier to locate, or providing touch screen facilities (these were not available for the trials, and were not evaluated). In addition, the microphone headset was not liked, and therapists suggested replacing it with a free-standing microphone (so long as guidance was given on the control of the microphone-to-mouth distance), or with individual lapel microphones for the therapist and client.

Within the individual modules, therapists suggested that the pitch games could have a wide band within which pitch has to be controlled; the band could be narrowed as the client's ability improves, to encourage better pitch control. There was also a request for more modules dealing with pitch reduction or lowering, along the lines of the collapsing building display. In the vowel modules, some facility for returning the moveable vowels to original positions was suggested, while in the /s/-/sh/ module the addition of the image of a face showing the changing lip-rounding, was proposed.

The final suggestions included offering the system different minority language versions such as Punjabi or Arabic (according to local conditions), and a facility for obtaining printouts of targets and/or users' attempts for classroom follow-up and homework use.

7.5 Conclusions

The results of the user trials were encouragingly positive, both about the therapists' assessment of their client's progress, and of the client's own experience of the system. In the therapists' assessments, the large majority of users appeared to enjoy using the system, even where they reported problems with some of the modules. Most, too, responded positively to the feedback they were given, and were keen to use the system again. Most therapists, however, did not think that clients were ready to use the system without supervision. All the user groups were similar in their responses, with the exception of the cochlear implant users, who experienced slightly higher levels of frustration and difficulty with the system. This may be because the system did not provide any auditory feedback, which is a central part of their therapy. In the clients' own assessments, the system scored highly for enjoyment, accessibility, ease of use, and interpretability, and most people, including the very young children, felt comfortable with it. Again, the cochlear implant users were slightly more reserved in their judgements than other groups. The accuracy of the speech analysis for all the real-time modules, resulted from the implementation and optimisation of appropriate algorithms on the PC platform (as presented in Chapter 6), was generally considered satisfactory. Visual feedback such as the *linegraph* as a feedback for pitch (the "pitch aeroplane"), the *vowel display,* and the fricative analysis module, were found highly motivating. The range of animated sequences showing consonants production were found to have great potential for helping clients to understand what was happening inside their mouth.

In summary, while there are a number of aspects in which the system can be improved, and several recommendations for change (as presented in Section 7.4), the results of the questionnaires confirmed the impression gathered throughout the user trials that this system is enjoyable to use, effective at stimulating users interest in their speech therapy, and capable of motivating them to try to improve their speech.

CHAPTER 8

Conclusions

Visual feedback techniques, together with other feedback methods such as tactile aids, are alternatives for providing the hearing impaired with the necessary information about their own speech, which is missing because of the hearing impairment. In the last twenty five years there have been many attempts to build effective visual feedback schemes for speech rehabilitation. However, only a few give some limited positive results. This thesis has analysed the causes of such failure, and proposed a novel approach for improved design of visual feedback for the hearing impaired. Chapter 2 detailed the necessary background information about the speech production process on both normal and hearing-impaired people. This process, explained using the Acoustical Theory of Speech Production, is the basis for understanding the speech analysis algorithms implemented in the last part of the thesis. Differences between normal speech and hearing-impaired speech were highlighted for both categories of hearing impairment (pre-lingual versus post-lingual deafness). Chapter 3 continued the literature survey, with an extensive review of previous work in the field of visual feedback techniques in the rehabilitation of hearing-impaired speech. Examples of screens from a variety of published systems were reported, and a critique of the design methodologies made, since it appears that visual feedback in published systems has been designed using a practical approach, instead of being the result of some user-oriented research.

The original work of the thesis starts with Chapter 4, where the problems that affect published systems were highlighted. A survey of twelve British therapists for the hearing impaired, teachers of the deaf, linguists and student speech therapists, reported problems such as negative reinforcement, lack of motivation, frustration, inflexibility, and difficulties in understanding how to correct speech productions. The possible reasons for these problems were discussed in the following areas: inadequate interface and system design, causing difficulties in setting and controlling the system; little research in visual / auditory perception causing difficulties in associating visual feedback with particular speech features; lack of evaluation of the aids in therapy clinics, causing systems to have problems such as negative reinforcement, frustration, and ineffectiveness; and inadequate accuracy of speech processing, causing negative reinforcement or slow response. Therefore, in Chapter 4 a novel approach to designing visual feedback for the rehabilitation of hearing-impaired speech was proposed. The approach is based on the integration of knowledge on visual interface design techniques, expertise of speech therapists, and results from new experiments on visual feedback responses for hearing-impaired speakers. Visual interface design techniques were presented, together with basic feedback control mechanisms in human behaviour, in order to construct a complete background on visual feedback design.

Chapter 5 covers the design and experimentation of visual stimuli on hearing-impaired and normal-hearing subjects. Three experiments were described. The goal of Experiment 1 was to test the response of subjects to these visual stimuli, in terms of intuitive links between visual dimensions and speech features. A set of appropriate visual stimuli were selected, covering both traditional techniques and new techniques never before used in published systems. These stimuli were presented to the subjects, who were asked to accompany the stimuli with their voice in a mode they considered appropriate. In this way an intuitive connection between visual stimuli and speech features was characterised. The results were evaluated with the help of three independent professional assessors who were expert phoneticians. In the evaluation of the results the following issues were investigated: interest / enjoyment, motivation to speak, significant changes in loudness, pitch and vowel quality. The experiment provided valuable information on intuitive connections between visual stimuli and speech features. Some of the stimuli already used as a feedback method in published systems were confirmed as effective, such as *dimension* and *length* as a feedback for loudness, and *line graph* as a feedback for pitch. New methods were also identified, such as *distance* and *luminance* for loudness, *speed* and *realistic stimuli* for pitch. All these modalities were considered for implementation in a prototype system. The results for adult speakers were quite different from those from children. Adults were collaborative and patient, and their behaviour was more consistent. Children were less consistent in their behaviour. They did tend to follow the stimuli with their voice for some of the stimuli, but the effect of these stimuli on their voice was not consistent.

The other two experiments investigated the use of alternative graphic techniques such as virtual reality and multimedia. Experiment 2 compared a traditional display technique (a computer screen) with recent 3D headsets. Hearing-impaired children tended to prefer the 3D option, which looks promising as a speech rehabilitation aid. Experiment 3 assessed the enjoyment of users presented with multimedia technology. Two short video-clips, selected with the help of speech therapists for their potential effectiveness in rehabilitating some problem in hearing-impaired speech, indicate that this methodology is motivating.

Chapter 6 gathers together the results of the experiments, the background knowledge on visual display design, and recommendations from the therapists, to design and implement a prototype system for the rehabilitation of hearing-impaired speech based on a set of visual feedbacks which try to improve the effectiveness of traditional techniques, and propose new methods. The speech processing algorithms were designed with the goal of fulfilling the requirements of accuracy and were implemented and optimised for real-time operation without the need of specialised hardware, reducing system cost and allowing portability of the system. Design and implementation was a cyclic process, done in close collaboration with speech therapists who tested the software at several stages of its development with hearing-impaired users, and recommended changes and improvements.

Chapter 7 describes the results of user trials conducted on the prototype system designed in Chapter 6 with pre-lingually, post-lingually, elderly and cochlear implanted hearing impaired. The user trials resulted in a set of recommendations for improving the system and in the general feeling that the system was enjoyable to use, stimulating, and capable of motivating the users to try to improve their speech.

Need for further research

The study on the design of visual feedback systems for hearing-impaired speech rehabilitation is relatively recent. Although this thesis has addressed some of the key issues for the implementation and development of such systems, many questions remain to be answered. For example, it is still unclear which visual feedback techniques are best suited for children. Further research in that area is required. While the display for vowels received a very positive response from therapists, further research into the intuitive links between different visual stimuli and vowel quality need to be carried out. Finally, further work is required on the beneficial effect of visual feedback to segmental aspects of speech which requires the use of more reliable models for consonants.

Appendix A

Questionnaires

I. Survey questionnaire

This questionnaire was designed to be used in semi-structured interwiew in order gather information from speech therapists about computer training aids.

1. Your background

1.1. What is your professional training / background?
- speech therapist
- hearing therapist
- teacher of the deaf
- other?

1.2. Do you have a hearing impairment yourself? If so, to what extent?
- moderate / profoundly?
- one ear / both ears?

1.3. Do you use a hearing aid of some sort?
- behind the ear
- cochlear implant
- radio microphone
- other

1.4. Do you teach spell skills (production or perception) to the hearing-impaired?
- what sort of problems do you deal with?

2. Computer training aids

2.1. Do you use / have you used a computer before?
- at home
- in the office
- at school

2.2. Do you enjoy using computers?
- yes
- use them if I have to
- no, dislike and avoid them
- never use them

2.3. Have you ever used a speech training aid based on a computer?

- IBM Speech Viewer
- C-speech
- Visispeech
- Visipitch
- other (name?)

2.4. What did you use it for?

2.5. Was it useful in your teaching?

- why / why not?

2.6. What did you like most about it?

- why?

2.7. What did you like least about it?

- why?

2.8. Have you ever used any other sort of aid, e.g. tactile aids, palatography?

- what sort of aid?
- what for?

3. Pronunciation teaching

3.1. What are your goals in teaching pronunciation to hearing-impaired speakers?

- complete naturalness
- complete intelligibility
- improved naturalness
- improved intelligibility
- improved self-confidence
- improved self-monitoring
- improved motivation
- other?

3.2. Do you give your speakers any sort of assessment of their speech skills?

- production
- perception
- before therapy
- during therapy (monitoring progress)
- after therapy

3.3. What forms of assessment do you use?

3.4. Do you use any standardised assessment procedures?

- PASC (Grunwell)
- PETAL (Parker)
- Vocal Profile Analysis (Laver)
- Edinburgh Articulation Test
- Nuffield Dyspraxia Programme
- other (please name)

3.5. Which areas of pronunciation do you feel are most important to concentrate on?

3.6. Which areas of pronunciation seem to come up most often as problems for your speakers?

3.7. How important is the speaker's own assessment of their needs in the success of therapy?

3.8. Do you give explicit articulatory information to your speakers?
- explicit instructions to move / place articulators
- explicit articulatory diagrams (e.g. vocal tract)
- information restricted to visible articulators (e.g. tongue and lips)

3.9. Do you use any published teaching materials?
- complete speech training courses?
- ad hoc examples from books?

3.10. Which materials do you use / prefer?
- (please name)

4. Courseware

4.1. Which sort of courseware do you think would be most useful?
- a set of flexible modules dealing with the main pronunciation problems, each independent of the others
- a set of modules to be taken in a particular sequence, forming a structured "curriculum" or teaching programme
- a set of basic teaching modules, with optional sets of prectice drills and exercises to follow
- something else?

4.2. Do you think it would be useful to have a system which could teach autonomously (i.e. a system which allowed "self-access")?
- yes
- no - teachers should always be present

4.3. What are your main reservations about such a system?
- ability to detect errors
- ability to diagnose / correct errors
- robustness
- possibility of reinforcing incorrect speech
- inability to explain
- speakers' computer-shyness or fear of technology
- cost
- other

4.4. Which groups of speakers do you think a system like this would be useful for?
- children
- adults
- cochlear implant patients
- elderly speakers
- any other group
- none of these groups

4.5. How could it be changed to make it more suitable for the group(s) which you deal with in your work? (please name them)

4.6. Which teaching functions should it perform?

- basic exposition of problems and how to correct them
- interactive diagnosis and teaching with feedback on performance
- routine drills and practice after lesson with teacher
- assessment of problems
- assessment / monitoring of progress during therapy

4.7. Would you like to see an "Introductory Module", teaching speakers how to use the system?

- helpful
- essential
- not needed

4.8. Should speakers learning one feature (e.g. intonation) be made aware of mistakes in other aspects of pronunciation (e.g. segmental)?

- yes, all the time
- no, this is demotivating
- gradually, not at first but maybe later on
- under the control of the teacher
- other (please explain)

4.9. What level of detail should the stress / intonation displays show?

- whole sentence / phrase / word only
- individual words
- syllables
- segments
- teacher should be able to choose

4.10. Would you like to be able to see more detailed data about clients' speech for your own diagnosis purposes?

- spectrograms
- fundamental frequency
- vowel formant frequencies

4.11. What sort of diaplay would you like to see?

4.12. Should there be a separate set of modules to teach basic control of speech features?

- fundamental frequency
- loudness
- voicing
- speech and segment duration
- pausing
- other

4.13. Would it be helpful to be able to review clients' past results in order to assess overall improvement?

4.14. Please name the 3 most important modules which could be offered by the system for the group of speakers you normally deal with. (*)

- pitch control
- intonation level / range
- linguistic use of intonation (statements vs questions, e.g.)
- voice quality (hashness, breathiness, falsetto)
- vowel contrasts
- consonant production
- word stress
- rhythm
- other

(*) Which group of speakers is this?

4.15. How would you prefer the lessons to be ordered?

- speakers work their way through a set of lessons in the order they appear
- speakers pick and choose the lessons they want to do each day
- teachers pick and choose the lessons they want to do each day

5. Feedback

5.1. How do you find the pictures used for feedback (in the system you use)?

- useful
- understandable
- effective
- not effective

5.2. How do you find the exercises?

- interesting
- easy to do
- difficult
- boring
- frustrating

5.3. When speakers' attempts are rejected by the system (ignoring for now whether they were actually acceptable to you as a listener), do the pictures explain well enough what was wrong?

5.4. From the pictures, would speakers generally know how to change their pronunciation to get it right?

5.5. Would you prefer more information on what was wrong?

5.6. What sort of information would be helpful?

5.7. Would it be helpful for speakers to get a "score" to show how well they have done?

5.8. Should something interesting happen when they get it right?

5.9. Where speakers discouraged when they got it wrong?

5.10. Is the response fast enough?

5.11. Would it help if the computer had some way of indicating when it was busy processing what the speaker had said?

5.12. How soon after speakers have spoken should the response be?

5.13. Would it be useful to see the display moving as they speak?

5.14. Would tactile feedback (vibration) be useful?

- for the examples
- for their own speech as they speak
- for their own speech as they have spoken (on playback)
- other

6. Control

6.1. Do you find easy / hard to control the system?

6.2. What do you find most difficult?

- pressing buttons with the mouse
- knowing where to look for buttons
- what each butto does
- which window to look at
- other

6.3. How would you like to control the computer?

- using keys
- using the mouse (point and click)

6.4. What other form of control would you prefer?

- a special separate key, maybe with just a few colour-coded keys?
- a remote control, like on the TV / video
- a touch sensitive screen

6.5. Would a TV / video remote control be a useful way of controlling the system?

- don't like using them
- yes

6.6. How would you rate the degree of control available to speakers at present?

- enough control: it is easy to move around when they want
- not enough control: they should have more choice in what to do and where to go
- too much control: too many buttons to click / decisions to make

6.7. In case you use a headset microphone, how do you find it?

- easy to use
- comfortable
- awkward
- hard to use
- uncomfortable

6.8. Can you foresee any groups of speakers for whom the headset microphone would be unsuitable?

6.9. Which type of microphone do you prefer?

- a microphone on the table
- a hand-help microphone
- a built-in microphone (part of the computer)
- a headset microphone

6.10. Are you happy with the way the system responds when speakers make a genuine mistake, e.g. by speaking at the wrong time or not saying the right words?

- yes, happy with the way it responds
- no, could be improved (in what way?)

6.11. Was it clear when to speak and when not to speak?

6.12. Do you always know what is happening, and what to do next?

6.13. Would you be happy with speakers using the system on their own?

- if not, why not?

7. Comments and suggestions

- (discussion)

II. Evaluation questionnaires

A set of evaluation questionnaires for the therapists and their clients were designed to answer general questions of usability and efficacy of the prototype system.

Three questionnaires were designed in total:

1. the therapist's evaluation of the client's performance with the system (a 10-item questionnaire)
2. the client's own assessment of the ease of use of the system (10 items)
3. children's own assessment of the ease of use of the system (5 items)

Therapist's evaluations of client's performance

This questionnaire was intended to be completed by the therapist immediately after a session with a client. It contained 10 statements, covering both positive and negative aspects of the use of the system, with which the therapist could agree or disagree, providing their views on a 5-point scale.

Q1 The client found this system enjoyable to use:

Q2 The client found this system boring to use

Q3 The client was keen to use the system again

Q4 The client found the system frustrating to use

Q5 The system helped the client's progress towards the goals set for their therapy

Q6 The system has made the client more keen to receive therapy

Q7 The client is ready to use the system by themselves

Q8 The client's confidence in his / her speech production improved as a result of using the system

Q9 The client responded positively to the feedback provided by the system

Q10 The system made their speech performance decline

Clients' evaluations of system usability

This questionnaire was intended to be given to the client after all their therapy sessions had been completed. It contained 10 statements, again covering both positive and negative aspects of the use of the system, with which they could agree or disagree using the same 5-point scale. This version of the questionnaire was designed for use by adults.

Q1 I think that I would like to use this system a lot

Q2 It took me too long to access parts of the system I needed to use

Q3 I thought the system was easy to use

Q4 I think I would always need the help of a therapist to be able to use this system

Q5 This system gives excellent coverage of the speech features I need to work on

Q6 I thought the pictures were easy to interpret

Q7 Most people would learn to use this system very quickly

Q8 I found it difficult to adapt the system to my needs

Q9 I felt flustered using the system

Q10 I needed to learn a lot of things before I could get going with the system

Children's evaluations of system usability

This questionnaire was intended to be given to the children after all their therapy sessions had been completed, to be completed with the help of the therapist if necessary. It contained five statements corresponding approximately to Questions 1, 4, 6, 7 and 9 of the adult version of the questionnaire, again covering both positive and negative aspects of the use of the system.

Q1 I think I would like to use this system a lot

Q2 I think I would need my therapist to help me

Q3 I knew what the pictures were telling me about my speech

Q4 I think you could learn to understand the pictures quickly

Q5 The system made me feel nervous

Appendix B

I. DSP modules

Data acquisition code on the DSP32C board

This code, written in assembler for the AT&T DSP32C, was originally used for audio data acquisition. Data were processed by the PC, or (in a variant of this program) by the same DSP using programs written in "C" language. More recently, all the code running on the DSP board was moved on the PC's CPU.

```
/*============================================================================
        DATA ACQUISITION

        Fabrizio Carraro, 26 October 94
        Centre for Communication Interface Research
        University of Edinburgh
============================================================================*/
/*
DSP32C code residing in the LSI DSP32C card. Acquires data at
80kHz via channel A of the ADC card and performs 63 point FIR filtering in
order to cut off components above 10kHz.
The resultant signal is downsampled to 20 kHz, and passed into a swinging
buffer system. The PC host is notified via mailbox (shared memory locations)
when each buffer becomes full. Protective software, to alert the user of PC-DSP
catchup is included.

*/

#define FULL    0x1111
#define EMPTY   0xaaaa

.=0x000000                      /* jump to start */
        goto START
        nop
mailbox:        int 0           /* general host/dsp comms box */

/* IMPORTANT: below registers MUST be in locations n0 & n2 (byte addressing).
This allows an XOR operation on the second LSB to perform a toggle operation */
.= 0x00010
bufA_mb:        int FULL
bufB_mb:        int FULL

.align 4
START:  nop
        r8 = 0x0333             /* prepare to disable all interrupts */
        r10e = mailbox          /* set up general mailbox pointer */
        r5e = 0x200008          /* ADC CONTROL reg address*/
        r6 = 65411              /*sample at 80 kHz */
        pcw = r8                /* STOP interrupts  NOW */
        r8 = 0x0733             /* prepare to enable ADC interrupt only */

        /* output data is held in one of two swinging buffers. The first
           is placed at 0x800000. To enable easy toggling between buffers,
           an XOR action is used. This effectively assumes the alternative
           buffer exits at the location corresponding to an inversion of the
           12th bit (0x800800) */

        r1e  = 0x800000         /* r1 = pointer in buffer */
        r2e  = 0x000800         /* buffer toggle switch */
        r3e  = 0x800000         /* r3 = start of current buffer */

        r4e  = 0x200000
        *r5  = r6               /* set for 20 kHz sampling */
                                /* r12 = buffer A (initial state) */
        r12 = FULL
        r7 = 0x03e7             /* BUFFER SIZE  - 50ms at 20kHz */
        r22e = IVTP             /* interrupt vector pointer set */
```

```
        r5e = bufA_mb           /* pointer to current buffer flag */
        *r10 = r12              /* let the host know its running */

        r13 = 0                 /* MAX value in frame (for VU meter) */
        r14 = EMPTY             /* empty buffer code */
        goto BEG                /* jump to start of main program */
        nop                     /* latent instruction */


        /* main program start ------------------------------------------*/
BEG:    r11 = *r10              /* examine mailbox */
        nop                     /* CAU register load has 1 cycle latency */
        r11 - 0x1010            /* Has PC-Host signalled to start? */
        if (ne) goto BEG        /* No, continue looping */
        r11e = newbuf
        pcw = r8                /* enable ADC interrupts */

GATHER: if (r7-- >= 0) goto GATHER /* wait 'til bufferful gathered */
        r7 = r7 + 1                /* latent instruction */

        r10e = mailbox          /* load reg. 10 with mailbox addr.*/
        r1e = r3^r2             /* fast switch of buffers */
        r7 = 0x03e7             /* set counter for 50ms again */
        *r5 = r12               /* flag full for current buffer */
        *r10 = r13              /* mailbox now holds max value of frame */

        r5e = r5^0x000002       /* Switch buffer flag pointer */
        r13 = 0                 /* reset max val. in frame */
        r8 = *r5                /* read current (new) buffer flag */
        r3e = r1                /* re-initialise buffer pointer */
        r8 - r14                /* compare with EMPTY flag */
        if (eq) goto GATHER     /* continue ONLY if new buffer empty */
        nop
        *r5 = r7                /* flag an error - any old value but FULL/EMPTY
        nop                        will do. PC must will pick it up imediately
                                */
HALT:   goto HALT
        nop

CATCH:  nop                      /* INTERRUPT CATCH */
        ireturn

FILL:   nop                      /* ADC interrupt catch */

        *r11++ = a1 = float(*r4) /* log value into new buffer */
        r9e  = data2             /* set up filter pointers */
        r11 - newbuf_end         /* only filter every 4th sample */
        if (ne) goto RETURN

        r6e  = coef
        r11 = temp
        r10e = data
        *r11 = a1 = int(a1)

        /* 63-tap FIR filter incorporating 4:1 down sampling */

        a0 = (*r10++ = *r9++) * *r6++

        ***** Note: repeat the following instruction 62 times:
        a0 = a0 + (*r10++ = *r9++) * *r6++

        r6 = *r11               /* retrieve max value */
        r11e = newbuf;          /* reset the newbuf pointer */
        *r1++ = a0 = int(a0)    /* store the final value in th eo/p buffer */
        r7 = r7 - 1             /* count the value */
        r13 - r6                /* compare current value with max so far */
        if (lt) r13 = r6        /* retain maximum */

RETURN: nop
        nop
        nop
        ireturn

/* -------------- interrupt vector table ----------------------------------*/
IVTP:   goto CATCH
        nop
        goto CATCH
        nop
        goto CATCH
        nop
        goto CATCH
        nop
        goto CATCH
        nop
        goto CATCH
        nop
        goto FILL
        nop
        nop;nop
        nop;nop
```

```
/*-------------- FIR filter coefs & data ---------------------------------*/
coef: float 0.000003, -0.000000, -0.000022, -0.000070
coef1: float -0.000122, -0.000124, 0.000000, 0.000289
coef2: float 0.000680, 0.000961, 0.000821, -0.000000
coef3: float -0.001477, -0.003129, -0.004041, -0.003185
coef4: float 0.000000, 0.005025, 0.010084, 0.012425
coef5: float 0.009416, -0.000000, -0.014051, -0.027807
coef6: float -0.034190, -0.026253, 0.000000, 0.043347
coef7: float 0.096689, 0.148521, 0.186213, 0.200000
coef8: float 0.186213, 0.148521, 0.096689, 0.043347
coef9: float 0.000000, -0.026253, -0.034190, -0.027807
coefa: float -0.014051, -0.000000, 0.009416, 0.012425
coefb: float 0.010084, 0.005025, 0.000000, -0.003185
coefc: float -0.004041, -0.003129, -0.001477, -0.000000
coefd: float 0.000821, 0.000961, 0.000680, 0.000289
coefe: float 0.000000, -0.000124, -0.000122, -0.000070
coeff: float -0.000022, -0.000000, 0.000003

data: 4*float 0.0
data2: 59 * float 0.0
newbuf: 4*float 0.0
newbuf_end:
temp:   int 0
```

Voiceless fricative analysis

The following function executes the "voiceless fricative analysis" based on a simplification of the method described by Wrench et al (1995). Requires the calculation of a FFT (not shown).

```
/************************************************
   gravityCentre

calculates the gravity centre of the spectrum
Doesn't take into account the beginning of the spectrum,
in order to avoid the effect of blowing noise
The array "y" contains the results of the FFT.
Fabrizio Carraro, CCIR
************************************************/
short gravityCentre(double *y, short npoints)
{
   #define Start 20
   short n, gc, err;
   double sumN = 0, sum = 0;
   double *dpy;

   struct {
      short gc;
      short loudness;
      short bandwidth;
      short optional1;
   } res;

   dpy = y+Start;
   for (n=0; n<npoints-Start; n++) sumN += *dpy++ * n;

   dpy = y+Start;
   for (n=0; n<npoints-Start; n++) sum += *dpy++;

   gc = (short) (Start + sumN/sum);

#if dll
   res.gc=gc;
   res.loudness = loudness;
   res.bandwidth = 1234;
   res.optional1 = 5678;
   err = PutRTConsonantTrackerFrame(&res) ;
#endif


   return(gc);

} /* gravityCentre */
```

Speech segmenter

The following code from the speech segmenter shows the section implementing the Viterbi algorithm. The code was optimised for speed for the PC Windows 3.1 platform (McInnes, Carraro, et al. 1992).

```
/**********************************
                Viterbi
**********************************/
void Viterbi (void)
{
   #define HOR 150

   int X; /* token */
   int Y; /* state */
   float ThisScore, DistanceM[MAX_APU], SelfLoopingCostM[MAX_APU];
   float DurationCostM[MAX_APU];
   int  Count;
   float OldScore[MAX_APU];
   int OldDuration[MAX_APU];
   char msgBox[128];

   /* set at maximum the Score array,  set at 0 the Duration array and
      initialize the SelfLoopingCostM array */
   for (Y=0; Y<=Napu; Y++) {
      Score[Y] = MAXFLOAT;
      Duration[Y] = 0;
      DistanceM[Y] = DurationCostM[Y] = 0;
      SelfLoopingCostM[Y] = SelfLoopingCost(Y);
   }

   /* initialize PrecScore and totalDistance (backward) */
   Score[0] = 0;
   totalDistance = MAXFLOAT;

   /* Start of outer loop */
   for (X=1; X<=Ntoken; X++) {                  /* for each token...*/

      /* First inner loop: shift old scores and durations */
      for (Y=Napu; Y>=0; Y--) {                 /* for each state... */
         OldScore[Y] = Score[Y];
         OldDuration[Y] = Duration[Y];
         DurationCostM[Y] = DurationCost(Y);
      }

      /* Second inner loop: DP recursion */
      for (Y=Napu; Y>=1; Y--) {                 /* for each state... */

         Duration[Y]++;

         /* calculate distance and costs once: */
         DistanceM[Y] = Distance(X,Y) * Apu[Y]->Weight;

         totalDistance = OldScore[Y] + SelfLoopingCostM[Y];

         /* Initialize with horizontal step */
         Back[X][Y] = HORIZONTAL;

         /* check max duration for horizontal path: */
         if (Duration[Y] >= Apu[Y]->MaxDuration) totalDistance = MAXFLOAT;

         /* path decision: check all predecessors looking for the best one */
         for (Count = 0; Count < NPrev[Y]; Count++) {
             ThisScore = OldScore[ PrevState[Y][Count] ] +
                DurationCostM[PrevState[Y][Count]];

             if (ThisScore < totalDistance) {
                totalDistance = ThisScore;
                Back[X][Y] = Count+1;
                Duration[Y] = 1;
             }
         }
         Score[Y] = totalDistance + DistanceM[Y];

      } /* end of column */

      Score[0] = MAXFLOAT;

   } /* End of outer loop */

   /* find final APU (predecessor of final NULL node): */
   X = Ntoken+1;
   Y = Napu+1;
```

```
   totalDistance = MAXFLOAT;
   for (Count = 0; Count < NPrev[Y]; Count++) {
       ThisScore = Score[ PrevState[Y][Count] ] + DurationCost(PrevState[Y][Count]);

       if (ThisScore < totalDistance) {
          totalDistance = ThisScore;
          Back[X][Y] = Count+1;
       }
   }

   /****** TraceBack *******/
   X = Ntoken+1;
   Y = Napu+1;
   BackTracing(X, Y);

   return;
} /* Viterbi */



/*****************************
   Distance
 Squared Euclidean distance
 - incorporating floating-point
 and integer arithmetic options
*****************************/
float Distance(int TokenPtr, int ApuPtr)
{
   /* vars used in the floating-point case */
   float *ptr1,*ptr2,*stop;
   float Dist;
   float DistTot = 0;

   /* vars used in the integer case */
   int *p1,*p2,*istop;
   int IDist;
   long IDistTot=0;

   if (integerDist) {   /* compute distance on integer coefficient values */

      p1 = &(iCentroid[ApuPtr].ipar[0]);
      p2 = &(iToken[TokenPtr].ipar[0]);
      istop = &(iCentroid[ApuPtr].ipar[CENTROID_PARMS-1]);

      while (p1 <= istop) {
         /* distance calculation */
         IDist = (*p1++)-(*p2++);
         IDistTot += IDist * IDist ;
      }
      return( (float)IDistTot/1024. );

   } else {     /* compute distance on floating-point values */

      ptr1 = &(Apu[ApuPtr]->Centroid[0]);
      ptr2 = &(Token[TokenPtr].Par[0]);

      stop = &(Apu[ApuPtr]->Centroid[CENTROID_PARMS-1]);

      while (ptr1 <= stop) {
         Dist = *(ptr1++) - *(ptr2++) ;
         DistTot += Dist * Dist ;
      }
      return(DistTot);

   }

} /* Distance */


/*********************************************
   SelfLoopingCost
Returns the cost of remaining in the same state
(has to be properly weighted)
*********************************************/
float SelfLoopingCost (int Status)
{
   float Cost;

   Cost = Apu[Status]->SuccessorProbability;
   return( Cost /** SelfLoopingCostWeight */);

} /* SelfLoopingCost */



/*************************************************************************
   DurationCost
```

```
Gives the duration cost when changing state. Uses the Duration array
(updated by the Dtw loop) that records how long each state was repeated.
The duration cost is taken from the duration list of each APU.
If the MaxDuration is 9999 this means that the model may have an infinite
duration and the DurationCost must be zero.
*************************************************************************/
float DurationCost (int State)
{
   /*int MaxDur;*/
   float Cost;

   if (State == 0) return(0.);

   /* if the model may have infinite duration return 0 */
   if (Apu[State]->MaxDuration == 9999)  return(0.0);
   if (Duration[State] > Apu[State]->MaxDuration) return(MAXFLOAT);

   Cost = Apu[State]->DurationList[ Duration[State]-1 ];
   return(Cost);
} /* DurationCost */



/***********************************
   SetBack
packs two four bits info into
a byte. Uses the Back[][] array
***********************************/
void SetBack (int X,
              int Y,
              int Info)
{
   int Yh;
   unsigned char Was, New;

   Yh = Y>>1;
   Was = Back[X][Yh];
   if (Y & 1) {         /* odd address */
      New = Was & 0xf;  /* take the low nibble only */
      New = New | (Info << 4);
   }
   else {
      New = Was & 0xf0;
      New = New | Info;
   }
   Back[X][Yh] = New;

} /*SetBack */



/***********************************
   ReadBack
reads the four bit info packed by
the function SetBack.
Uses the Back[][] array.
***********************************/
int ReadBack (int X,
              int Y)
{
   unsigned char Both;

   Both = Back[X][Y>>1];
   if (Y & 1) return(Both >> 4);
   else return(Both & 0xf);

} /* ReadBack */



/***********************************
 Backtracking:
***********************************/
void BackTracing(int Xnd,
                 int Ynd)
{
   char BackContents;
char msgBox[128];

   Norm = 0;
   BackPtr = 0;
   while (Xnd != 0) {
      BackContents = Back[Xnd][Ynd];

      if (BackContents == HORIZONTAL) BackTra[BackPtr++] = Ynd;
      else Ynd = BackTra[BackPtr++] = PrevState[ Ynd ][ BackContents-1 ];
      Norm++;
      Xnd--;
```

```
    }
    Norm-=2;

} /* BackTracing */


/*******************************
   AnalyzeBestPath
*******************************/
void AnalyzeBestPath (void)
{
   int n, apuIndex, selfLoopCnt, tokIndex, previousApu, exclusiveFrames;
   float dis, slCost, DCost, AccumulatedDistance, MeanDistance;
   FILE *ShortResultFile, *DiffFile;

   ShortResultFile      = fopen(resultFileName,"wt");

   apuIndex = 1;
   selfLoopCnt = 0;
   tokIndex = 1;
   AccumulatedDistance = 0;
   MeanDistance = 0;
   exclusiveFrames = Norm + 1;

   previousApu = 9999;
   for (n = Norm; n >= 0; n--) {

      if ( (apuIndex = BackTra[n]) == previousApu) {
         selfLoopCnt++;
         dis = Distance(tokIndex,apuIndex) * Apu[apuIndex]->Weight;
         slCost = SelfLoopingCost(apuIndex);
         AccumulatedDistance += (dis + slCost);
         MeanDistance += dis;
      }
      else {                                    /* some diagonal path */
         if (previousApu == 9999) DCost = 0.;
         else {
            DCost = DurationCost(previousApu);
            if (Diff) fprintf(ShortResultFile, "%d %d %s %1.3f\n",
                              tokIndex-selfLoopCnt-1, tokIndex-1,
                              Apu[previousApu]->Name, MeanDistance/selfLoopCnt);
            else fprintf (ShortResultFile, "%d %d %s\n",
                     tokIndex-selfLoopCnt-1, tokIndex-1, Apu[previousApu]->Name);

            Duration[previousApu] = selfLoopCnt;

            /* if silence then remove its contribution to the total error */
            if (!strncmp(Apu[previousApu]->Name,"##",2)) {
               totalDistance -= MeanDistance;
               exclusiveFrames -= selfLoopCnt;
            }
         }
         apuIndex = BackTra[n];
         dis = Distance(tokIndex,apuIndex) * Apu[apuIndex]->Weight;
         MeanDistance = dis;
         AccumulatedDistance += dis + DCost;
         selfLoopCnt = 1;
      }
      previousApu = BackTra[n];
      tokIndex++;
   }

   if (Diff)  fprintf (ShortResultFile, "%d %d %s %1.3f\n",
                       tokIndex-selfLoopCnt-1, tokIndex-1, Apu[previousApu]->Name,
                       MeanDistance/selfLoopCnt);
   else    fprintf(ShortResultFile, "%d %d %s\n",
           tokIndex-selfLoopCnt-1, tokIndex-1, Apu[previousApu]->Name);

   //if (Diff) fclose(DiffFile);
   fclose (ShortResultFile);

} /* AnalyzeBestPath */
```

II. Graphics

Fast spectrogram display

The following code ("C" for Windows 3.1 or WinNT) resulted particularly fast in displaying the real-time spectrogram used in the "therapist's graphical output" of the vowel analysis. The LPC spectra calculation, used in the display, is not shown.

```
/*****************
   drawSpectrogram
*****************/
void drawSpectrogram(HWND hWnd, double *spect)
{
   HDC hdc;
   HPEN pen, oldPen;
   static int x=3;
   int y, shade, shift;
   double max,min,app, range;

   if (IsIconic(hWnd)) return;

   if (FFTORDER==8) shift = 1;
   if (FFTORDER==9) shift = 2;

   hdc = GetDC(hWnd);


   /* scan the spectrum to find the max and min */
   if (spectAgc) {
      max = -999999;
      min = 999999;
      for (y=18; y<FFTPOINTS/2; y++) {
         if ((app=spect[y<<1]) > max) max = app;
         else if (app < min) min = app;
      }
      min = log10(min+0.1);
      max = log10(max+0.1);
   }


   /* increment x each time */
   x+=3; if (x>510) x=1;

   /* for each dot */

   for (y=1; y<128; y++) {

      MoveTo(hdc,x,128-y);

      /* select color */
      if (!spectAgc) {
         shade = -20 +(int)(4*log10( spect[y<<shift]+0.1)) ;
      }
      else {
         range = (max - min) /16;
         shade = (int)( -min/(range) + log10( spect[y<<shift]+0.1) / (range)  );
      }

      if (shade>15) shade = 15;
      if (shade<0) shade = 0;
      oldPen = SelectObject(hdc, sPen[shade]);
      LineTo(hdc,x+3,128-y);
   }

   /* release the pen */
   pen = GetStockObject(BLACK_PEN);
   oldPen = SelectObject(hdc,pen);
   ReleaseDC(hWnd,hdc);

} /* drawSpectrogram */
```

Vowels display

The graphical module of the vowels display was implemented in MS-Visual Basic 3. The "movable clouds" were implemented as icons, since this solved the problems of restrictions in overlapping moving objects in VB 3. Each cloud consisted of several "transparent" icons linked together. In order to overlap the pointer, the library function "BitBlt" was used. The following subroutine contains the code to move the pointer over the clouds.

```
Sub movePointer (pixX As Integer, pixY As Integer)
   Static OldPixX As Integer, OldPixY As Integer

   'Following are static only to improve speed
   Static NewX As Integer, NewY As Integer, temp As Integer
   Static hDC As Integer, releaseit As Integer

   ScaleMode = 3 'PIXELS
   hDC = Me.hDC

   'put back old background but not the first time
   If started <> 0 Then
      temp = BitBlt(hDC, OldPixX, OldPixY, 32, 32, picCopy.hDC, 0, 0, SRCCOPY)
   End If
   'save new background
   temp = BitBlt(picCopy.hDC, 0, 0, 32, 32, hDC, pixX, pixY, SRCCOPY)
   'paint mask
   temp = BitBlt(hDC, pixX, pixY, 32, 32, PicMask.hDC, 0, 0, SRCAND)
   'paint sprite
   temp = BitBlt(hDC, pixX, pixY, 32, 32, PicSprite.hDC, 0, 0, SRCPAINT)
   started = 1

   'save old coords
   OldPixX = pixX: OldPixY = pixY
   ScaleMode = 1 'TWIP

End Sub
```

Changing an object's size

Changing an object's size, such as in the case of the "Dimension" feedback for loudness ("Dog & bone") and the "Distance" feedback for loudness ("Talking face"), is straightforward in Visual Basic. The "move method" is a library function call which accepts as an optional parameters the size of the object being moved. The following code is used in the "Distance" feedback, for changing the aspect and dimension of the character, following the speaker's loudness. The code includes also a simple IIR low-pass filter used to smooth the display visual response.

```
Sub tmrGame2_Timer ()

    Static Fault As Integer
    Static YInc As Long       'Current Delata Amp Value/Image Position
    Static loudness As Long, OldLoudness As Long
    Static Buf(2) As Integer
    Dim inc, hShift, vShift As Integer

    ' read loudness
    Fault = GetRTPitchTrackerFrame(PitchResults(0))

    ampvalue = Sqr(PitchResults(1)) * hScroll1.Value

    'low pass filter
    If check1.Value Then
        loudness = (OldLoudness * .8) + (ampvalue * .2)
    Else
        loudness = ampvalue
    End If

    OldLoudness = loudness

    face = 0
    If loudness > 700 Then face = 1
    If loudness > 1100 Then face = 2
    If loudness > 2000 Then face = 3
    If loudness > 2700 Then face = 4
```

```
    If face <> oldFace Then
        imgFace(oldFace).Visible = False
        imgFace(face).Visible = True
        labelComment(oldFace).Visible = False
        labelComment(face).Visible = True

    End If

    ' resize face
    imgFace(face).Move 200, 200, loudness / 11, loudness / 11

    bigauge1.Value = loudness

    oldFace = face

End Sub
```

Bibliography

A

Angelocci, A. (1962), "Some observations on the speech of the deaf", *Volta Review,* vol. 64, pp. 403-405.

Angelocci, A., Kopp, G. and Holbrook, A. (1964b), "The vowel formants of deaf and normal hearing eleven to fourteen year-old boys", *Journal of Speech and Hearing Disorders,* vol. 29, pp. 156-170.

Apple Computer Inc (1993), "Macintosh Human Interface Guidelines", Reading, MA. Addison-Wesley Publishing Company.

Arends, N., Povel, D.J., Os, E. Van, Michielsen, S., Claasen, J., and Feiter, I. (1991) "An evaluation of the Visual Speech Apparatus", *Speech Communication,* vol. 10, pp. 405-414.

Arends, N. (1993), "The Visual Speech Apparatus", Instituut voor Doven, The Netherlands.

B

Bagshaw, P.C., Hiller, S.M., and Jack, M.A. (1993), "Enhanced pitch tracking and the processing of F0 contours for computer aided intonation teaching", *proceedings of Eurospeech 93*, vol.2, pp-1003-1006.

Baken, R.J. (1987), "Clinical Measurement of Speech and Voice", College-Hill Press, Boston.

Bailey, R.W. (1982), "Human performance engineering: A guide for system designers", Englewood Cliffs, NJ: Prentice-Hall.

Bass, L. (1993), "Architecture for interactive software systems: rationale and design" In: Bass, L. & Dewan, P. (Eds.), *User interface software*, New York, John Wiley & Sons; pp. 31-44.

Benbasat, I., Dexter, A.S., and Todd, P. (1987), "The influence of color and graphical inforation presentation in a managerial decision simulation", *Human-Computer Interaction,* 2, 65-92.

Berg, J.W. (Van Den) (1954), "Mechanism of the Larynx and Laryngeal vibrations", *Manual of Phonetic*, cap. 9, pp. 278-309, North Holland Co, Amsterdam.

Bernstein, J. (1989), "Application of speech recognition technology in rehabilitation", in *Speech today and tomorrow: proceedings of a conference at Gallaudet University, September 1988,* ed. B.M. Virvan, pp. 181-187, Gallaudet University, Washington.

Bernstein, L. E. (1989), "Computer Based Speech Training for Profoundly Hearing Impaired Children: Some Design Considerations", *The Volta Review,* 1989.

Bernstein, L. E. (1995), "Toward future tactile aids", In: Plant G, Spens K-E, eds. *Profound Deafness and Speech Communication.* London: Whurr Publications, 1995.

Bernstein, L. E. Demorest, M.E., Coulter, D.C. (1992), "Lipreading sentences with vibrotactile vocoders: Performance of normal-hearing and hearing-impaired subjects", *Journal of the Acoustical Society of America, 1992 Mar, v91*

Binnie, C., Daniloff, R., and Buckingham, H. (1982), "Phonetic disintegration in a five-year-old following sudden hearing loss", *Journal of Speech and Hearing Disorders,* vol. 47, pp. 181-189.

Boone, D. (1966), "Modification of the voices of deaf children", *Volta Review,* vol. 68, pp. 686-692.

Boothroyd, A., Nickerson, R.S., and Stevens, K.N. (1974), "Temporal patterns in the speech of the deaf: an experiment in remedial training", SARP 15, Northampton MA Clarke School for the Deaf, Research Dept.

Bot, K. de (1983), "Visual feedback of intonation I: Effectiveness and induced practice behaviour", *Language and Speech,* 26, pp. 331-350.

Burns, M.J., Warren, D.L, and Rudisill, M. (1986), "Formatting space-related displays to optimise expert and non-expert user performance", *Proceedings of CHI '86 Conference on Human Factors in Computing Systems,* Boston, Mass, April 1986, New York, Association for Computing Machinery.

Bryson, S. (1995), "Approaches to the Successful Design and Implementation of VR Applications", Computer Science Corporation / NASA Ames Research Center, Mottfett Field, Ca.

Brooks, S., Fallside, F., Gulian, E., and Hinds, P. (1981), "Teaching vowel articulation with the computer vowel trainer: Methodology and results", *British Journal of Audiology,* vol. 15, pp. 151-163.

Brou, P., Sciascia, R.T., Linden, L., and Lettvin, J.Y. (1986), "The Colors of Things", *Scientific American,* Vol. 255, n.3, September 1986.

Burroughs Corporation (1986), "InterPro (TM) user interface standards", *Human Factors Section,* System Products Group, Unisys Corp., 19 Morgan, Irvine, CA.

Bush, M. (1981), "Vowel articulation and laryngeal control in the speech of the deaf", unpublished doctoral dissertation, Massachussetts Institute of Technology.

C

Callan, J.R., Cullan, L.E., and Lane, J.L. (1977), "Visual search times for Navy tactical information displays (Report # NPRDC-TR-77-32). San Diego, CA. Navy Personnel Research and Development Center.

Calvert, D.R. (1961), "Some acoustic characteristics of the speech of profoundly deaf individuals", unpublished doctoral dissertation, Massachusetts Institute of Technology.

Calvert, D.R. (1962), "Deaf voice quality - a preliminary investigation", *Volta Review,* vol. 62, pp. 402-403.

Card, S.K. (1982), "User perceptual mechanism in the search of computer command menus", Proceedings: *Human Factors in Computer Systems* (Gaithersburg, MD). New York, Association for Computing Machinery.

Card, S.K. (1984), "Human Limits and the VDT Computer Interface", in Bennet, Jphn, Case, Donald, Sandelin, Jon, and Smith, Michael (eds.), *Visual display terminals: Usability issues and Health Concerns*, Englewood Cliffs, N.J.

Card, S. K., Moran, T. P., & Newell, A. (1983), "The Psychology of Human-Computer Interaction", Hillsdale, NJ: Erlbaum.

Champness, B.G., and Ikhlef, A. (1982), "Subjective reactions and performance of teletext viewers in response to graphics, coloured text", Technical report, Alternate Media center, New York University.

Christie, B. (1985) (Ed), "Human factors of information technology in the office", Chichester, John Wiley and Sons.

Cole, R.A., and Zue, V.W., (1979), "Speech as the eye sees it", In R.S. Nickerson (Ed.), *Attention and performance,* Hillsdale, NJ: Erlbaum.

Coleman, R.F., Mabis, J.H., and Hinson, J.K. (1977), "Fundamental frequency - Sound pressure level profiles of adult male and female voices", Journal of Speech and Hearing research, vol. 20, pp. 197-204.

Colton, R., and Cooker, H.S. (1968), "Perceived nasality in the speech of the deaf", *J. Speech and Hearing Research*, vol. 11, pp. 553-559.

Cowie, R., and Douglas-Cowie, E. (1983), "Speech production in profound post-lingual deafness", in *Hearing Science and Hearing Disorders,* ed. M.E. Lutman and M.P. Haggard, Academic Press, London, pp. 183-230.

D

Danchak, M.M. (1976), "CRT displays for power plants", *Instrumentation Technology,* 23(10), 29-36.

Delattre, P. (1954), "Les attributs acoustiques de la nasalité vocalique et consonantique", *Studia Linguistica,* VIII, vol. 2, pp. 103-109, reprint in *Studies in French and Comparative Phonetics,*London-Paris, 1966.

Delattre, P. (1965), "La nasalité en français et en englais", *The French Review,* vol. 39, pp. 92-109, reprint in *Studies in French and Comparative Phonetics,*London-Paris, 1966.

Denes, P.B. and Pinson, E.N. (1963), "The Speech Chain", Bell Telephone Laboratories, Inc., New Jersey. (Source: Furui)

Dodson, D.W., and Shields, N.L., Jr. (1978), "Development of user guidelines for ECAS display design (Vol. 1) (Report No. NASA-CR-150877), Huntsville, AL, Essex Corp.

Duchnicky, R.L., and Kolers, P.A. (1983), "Readability of text scrolled on visual display terminals as a function of window size", *Human Factors*, 25, 683-692.

Dunn, H.K. (1961), "Methods of Measuring Vowel Formant Bandwidths", J.A.S.A. Vol. 33, pp. 1737-1746.

E

Ellis, W., and Miller, D. (1981), "Left and wrong in adverts: neuropsychological correlates of aesthetic preference", *British Journal of Psychology*, 72, pp. 225-229.

Engel, S.E., and Granada, R.E. (1975), "Guidelines for man/display interfaces", (Technical report TR00.2720). Poughkeepsie, NY: IBM..

F

Faaborg-Andersen, K. (1957), "Electromyographic Investigation of Intrinsic Laryngeal Muscles in Human", Copenhagen.

Fallside, F, and Woods, W.A. (1985), "Computer Speech Processing", Prentice-Hall International UK.

Fant, G. (1956), "On the Predictability of Formant Levels and Spectrum Envelops from Formant Frequencies", *Readings in Acoustic Phonetics*, pp. 44-56, Rist. In *For Roman Jakobson,* Mouton, Le Hague, pp. 109-120.

Fant, G. (1960), "Acoustic Theory of Speech Production", Mouton's Co, Gravenhague.

Ferguson, J.B., Bernstein, L.E., and Goldstein, M.H. (1988), "Speech training aids for hearing-impaired individuals, II: Configuration of the Johns Hopkins aids", *Journal of Rehabilitation Research and Development,* 25.

Flanagan, J.L. (1965), "Speech Analysis, Synthesis and Perception", Springer-Verlag, Berlin Heidelberg, New York.

Flanagan, J.L. (1972), "Speech Analysis Synthesis and Perception", Springer-Verlag, Berlin

Flege, J.E., Fletcher, S.G., and Homedian, A. (1988), "An electropalatographic study of articulatory compensation in bite block produced /s/ and /t/", *Journal of the Acoustical Society of America,* 83, pp. 212-228.

Fletcher, H. (1934), "Loudness, pitch and the timbre of musical tones and their relation to the intensity, the frequency, and the overtone structure", *Journal of the Acoustical Society of America,* 6 (1934), pp. 59-69.

Fletcher, H. (1953), "Speech and Hearing in Communication", New York: van Nostrand, 1953.

Fletcher, S.G., Hasegawa, A. (1983), "Speech modification by a deaf child through dynamic orometric modelling and feedback", *Asha,* 48, pp. 178-185.

Fletcher, S.G., McCutcheon, M.J., Martin, J., and Smith, H.W. (1988), "Optoelectronic tongue height measurement in vowel production and modification". Submitted to the *Journal of Speech and Hearing Research.*

Fletcher, S.G., Dagenais, P.A., Critz-Crosby, P., (1991), "Teaching Vowels to Profoundly Hearing-Impaired Speakers Using Glossometry", the *Journal of Speech and Hearing Research, vol. 34, pp. 943-956.*

Foley, J.D., and Van Dam, A. (1982), "Fundamentals of Interactive Computer Graphics". Reading, MA, Addison-Wesley.

Forner, L., and Hixton, T.J., (1977), "Respiratory kinematics in profoundly hearing-impaired speakers", *Journal of Speech and Hearing Research,* vol. 66, pp. 373-408.

Fujimura, O., (1962), "Analysis of nasal consonants" *The Journal Of The Acoustical Society Of America*, 34(12), pp. 1865-1875.

Furui, S. (1989), "Digital speech processing, synthesis, and recognition", Marcel Dekker Inc., New York.

G

Galitz, W.O. (1989), "Handbook of Screen Format Design" (third edition). Wellesley Hills, MA: QED Information Sciences.

Garner, W.R., and Felfoldy, G.L. (1970), "Integrality of Stimulus Dimensions in Various Types of Information Processing", *Cognitive Psychology,* 1, 225-241.

Gay, T., Lindblom, B., and Lubker, J. (1981), "Production of bite-block vowels: Acoustic equivalence by selective compensation", *Journal of the Acoustical Society of America,* 69, pp. 802-810.

Geffner, D. (1980), "Feature characteristics of spontaneous speech production in young deaf children", *J Communication Disorders,* vol. 13, pp. 443-454.

Geffner, D. and Freeman, L. (1980), "Speech assessment at the primary level: interpretation relative to speech training", in *Speech assessment and speech improvement for the hearing impaired,* ed. J. Subtelny, A.G. Bell Assn. for the Deaf, Washington, DC.

Gilbert, H., and Campbell, M. (1980), "Speaking fundamental frequency in three groups of hearing impaired individuals", *Journal of Communication Disorders*, vol. 13, pp. 195-205.

Gold, B., and Rabiner, L. (1969), "Parallel processing techniques for estimating pitch periods of speech in the time domain", *Journal of the Acoustical Society of America,* 46 (2, part 2), pp. 442-448.

Goldstein, J. L. (1973), "An optimum processor for the central formation of pitch of complex tones", *Journal of the Acoustic Society of America*, 54 (1973), pp. 1496-1516.

Green, D.S. (1956), "Fundamental frequency of the speech of profoundly deaf individuals", unpublished doctoral dissertation, Purdue University.

Grether, W.F., and Backer, C.A. (1972), "Visual Presentation of Information". In van Cott, H.P., and Kinkade, R.G. (Eds) "*Human Engineering Guide to Equipment Design*", U.S. Government Printing Office, Washington D.C.

H

Harnad, S. R. (1987), "Categorical Perception", Cambridge University Press, 1987.

Harrington, J., (1994), "The contribution of the murmur and vowel to the place of articulation distinction in nasal consonants", *Journal of the Acoustical Society of America*, 96 (1), pp. 19-32.

Harrington, J., Laver, J. and Cutting, D (1986), "Word-structure reduction rules in automatic continuous speech recognition", *Proceedings of the Institute of Acoustics*, vol. 8 part 7, pp. 451-459.

Heider, F., Heider, G. and Sykes, J. (1941) "A study of the spontaneous vocalisations of fourteen deaf children", *Volta Rewiew,* vol. 43, pp. 10-14.

Heinz, J.M., and Stevens, K.N. (1961) " On the property of voiceless fricative consonants", *J. Acoust. Soc. Amer.* vol. 33, pag. 589-596.

Hess, W.H. (1983), " Pitch Determination of Speech Signals", Springer-Verlag, Berlin.

Hochberg, J. (1972), "The representation of things and people", in E.H. Gombricj, J. Hochberg & M. Black (eds.) *Art, perception and reality*, Baltimore, Johns Hopkins University Press.

House, A.S. and Stevens, K. (1956), "Analog Studies of the Nasalisation of Vowels", *Journal of Speech and Hearing Disorders*, vol. 21, pp. 218-232.

Hudgins, C.V. (1934),"A comparative study of the speech co-ordination of deaf and normal subjects", *J Genetic Psychology,* vol. 44, pp. 1-48.

Hudgins, C.V. and Numbers, F.C. (1942), "An investigation on the intelligibility of the speech of the deaf", *Genetic Psychology Monographs*, vol.25, pp. 289-392.

I

IBM (1993), "Object-Oriented Interface Design: IBM Common User Access Guidelines", IBM document SC34-4399, Carmel, IN, Que Publishing.

International Phonetic Association (1993), "Revision of the IPA", *Journal of the International Phonetic Association*, vol. 23 no. 1 (1993), pp. 32-34.
vol. 23 no. 1 (1993).

Irii, H., Itoh, K., and Kitawaki, N. (1987), "Multi-lingual speech database for speech quality measurements and its statistic characteristics", Trans. Committee on Speech Research, Acoust. Soc. Jap., S87-69.

J

Javkin, H., Antonanzas-Barroso, N., Das, A., Zerkle, D., Yamada, Y., Murata, N., Levitt, H., and Youdelman, K. (1993), "A motivation-sustaining articulatory/acoustic speech training system for profoundly deaf children", *Proc. IEEE ICASSP-93,* pp. 145-148.

K

Keister, R.S., and Gallaway, G.R., (1983) "Making software user friendly: An assessment of data entry performance". *Proceedings of the Human Factors Society 27th Annual Meeting,* 1031-1034. Santa Monica, CA: Human Factors Society.

Kelley, C.R. (1968), "Manual and automatic control", New York, Wiley.

Kemler Nelson, D.G. (1993), "Processing Integral Dimensions: The Whole View", *Journal of Experimental Psychology: Human Perception and Performance,* 1993, Vol. 19, No. 5, 1105-1113.

Kempelen, W.V. (1791) "Le mécanisme de la parole suivi de la description d'une machine parlante", J.V. Degen, Wien.

Kodak (1996), "Kodak Color Management Solution for Commercial Labs Delivers Consistent Color Throughout Workflow", http://www.kodak.com./aboutKodak/pressReleases/pr19970220-21.shtml

Kolers, P.A., Duchnicky, R.L., and Fergus, D.C. (1981), "Eye movement measurement of readability of CRT displays", *Human Factors*, 23, 517-527.

Krazenstein, C.G. (1782), "Sur la naissance et la formation des voyelles", *J. Phis*, Vol. 21 pp. 358-380.

Kurowski, K., and Blumstein, S., (1984), "Perceptual integration of the murmur and formant transitions for place of articulation in nasal consonants", *The Journal Of The Acoustical Society Of America*, 76 (2), pp. 383-390.

L

Ladefoged, P. (1982), "A Course in Phonetics", Harcourt Brace Jovanovich, publishers, San Diego, New York.

LaLomia, M.J., and Coovert, M.D. (1987), "A comparison of tabular and graphical displays in four problem solving domains", Unpublished technical report, Department of Psychology, University of South Florida, Tampa.

Lane, H. (1988), "Speech deterioration in postlingually deafened adults", unpublished manuscript.

Lane, H., and Webster, J.W. (1991), "Speech deterioration in postlingually deafened adults", *J. Acoust. Soc. Amer.*, vol. 89, pp.859-866.

Laver, J. (1980), "The phonetic description of voice quality", Cambridge University Press, Cambridge.

Laver, J., Wirz, S.L., Mackenzie, J., and Hiller, S. (1981), "A perceptual protocol for the analysis of vocal profiles", *Work in Progress, Department of Linguistics, University of Edinburgh,* vol. 14.

Leder, S., Spitzer, J., and Kirchner, J.C. (1987a), "Speaking fundamental frequency of postlingually profoundly deaf adult men", *Ann. Otol. Rhinol. Laryngol.*, vol. 96, pp. 322-324.

Leder, S., Spitzer, J., Kirchner, J.C., Flevaris-Phillips, C., Kirchner, J.C., and Richardson, F. (1987b), "Speaking rate of adventitiously deaf male cochlear implant candidates", *J. Acoust. Soc. Amer.*, vol. 82, pp. 843-846.

Leder, S., Spitzer, J., Milner, P., Flevaris-Phillips, C., Kirchner, J.C., and Richardson, F. (1987c), "Voice intensity of prospective cochlear implant candidates and normal hearing adult males", *Laryngoscop.*, vol. 97, pp. 224-227.

Levitt, H. and Nye, P.V. (Eds.) (1971), "Sensory training aids for Hearing Impaired", Washington D.C., National Academy of Engineering.

Levitt, H., Smith, C.R., and Stromberg, H. (1976), "Acoustical, articulatory and perceptual characteristics of the speech of deaf children", in *Proc of the Speech Communication Seminar,* ed. G. Fant, pp. 129-139, Wiley, New York.

Ling, D. (1976), "Speech and the hearing impaired child: Theory and practice", The Alexander Graham Bell Association for the Deaf Inc, Washington DC.

Lippmann, R.P. (1985), "A Review of Research on Speech Training Aids for the Deaf", in *Speech and Language: advances in basic research and practice*, vol. 7, ed. N.J. Lass, pp.105-133.

Lippmann, R.P., and Watson, C.S. (1979), " New computer-based speech training aid for the deaf", *J. Acoust. Soc. Amer.*, vol. 49, pp. 467-477.

Lockheed Missiles and Space Company (1983), "Human factors engineering standards for information processing systems", Sunnyvale, CA.

Mahshie, J.J. (1980), "Laryngeal behaviour in hearing-impaired speakers", unpublished doctoral dissertation, Syracuse University.

McCracken, W, and Sutherland, H. (1991), "Deaf-Ability, Not Disability: A Guide for the Parents of Hearing Impaired Children", Multilingual Matters LTD, Clevedon, Philadelphia, Adelaide.

Mahshie, J.J., Bernstein, L.E., Vari-Alquist, D., Waddy-Smith, B. (1988) "Speech training aids for hearing-impaired individuals: III. Preliminary observations in the clinic and children's home", *Journal of Rehabilitation Research and Development*, vol. 25 No. 4, pages 69-82.

Maki, D. (1983), "Application of the speech spectrographic display in developing articulatory skills in hearing impaired adults", in: I. Hochberg, H. Levitt & M.J. Osberger (Eds.) *Speech of the Hearing Impaired: Research, Training, and Personnel Preparation,* University Park Press, Baltimore.

Malone, T.W. (1981), "Toward a theory of intrinsically motivating instruction", *Cognitive Science,* 4, pp. 333-369.

Malone, T., and Lepper, M. (1983), "Motivation, cognition, and computerised instruction". Paper presented at the Office Naval Research Conference on Aptitude, Learning, and Instruction: Conative and Affective Process Analysis, Stanford, CA.

Marcus, A. (1992), "Graphic Design for Electronic Documents and User Interfaces", New York, ACM Press.

Marcus, A. (1995), "The Cross-GUI Handbook for Multi-platform User Interface Design", Addison Wesley Publishing, 1995.

Markel, J.D., (1972), "Digital Inverse Filtering, A New Tool for Formant Trajectory Estimation", IEEE Trans. AU-20, pp. 129-137.

Markel, J.D., and Gray, A.H., (1976), "Linear Prediction of Speech", Berlin, Springer-Verlag, pp.157-158.

Markides, A. (1970), "The speech of deaf and partially-hearing children with special reference to factors affecting intelligibility", *British Journal of Disorders of Communication*, vol.5, pp. 126-140.

Martony, J. (1965), "Studies on the speech of the deaf", Quarterly Progress and Status Report, Speech Transmission Lab, Royal Institute of Technology, Stockholm.

Martony, J. (1968), "On correction of voice pitch level for severely hard-of-hearing subjects", *American Annals of the Deaf,* vol. 113, pp. 195-202.

McGarr , N.S. and Harris, K.S. (1980), "Articulatory control in a deaf speaker", *Haskins Laboratories Status Report on Speech Research,* vol. SR-63/64, pp. 309-332.

McGarr , N.S. and Løfqvist, A.(1982), "Obstruent production in hearing-impaired speakers: interarticulator timing and acoustics", *J. Acoust. Soc. Amer.,* vol. 72, pp. 34-42.

McGarr , N.S. and Osberger, M.J. (1978), "Pitch deviancy and the intelligibility of deaf children's speech", *J. Communication Disorders.,* vol. 11, pp. 237-247.

McInnes, F.R., Carraro, F., Hiller, S.M., and Rooney, E.J. (1992), "Evaluation and Optimisation of a Segmenter for a PC-Based Pronunciation Teaching System", *Proc. Institute of Acoustics*, 14, pp. 109-116.

Medan, Y., Yair, E., and Chazan, D. (1991), "Super resolution pitch determination of speech signals", *IEEE Trans. Signal Processing,* ASSP-39 (1), pp. 40-48.

Melara, R.D. (1989), "Dimensional Interaction Between Color and Pitch", *Journal of Experimental Psychology: Human Perception and Performance,* 1989, Vol. 15, No. 1, 69-79.

Melara, R.D., and Marks, L.E. (1990), "Perceptual Primacy of Dimensions: Support for a Model of Dimensional Interaction", *Journal of Experimental Psychology: Human Perception and Performance,* 1990, Vol. 16, No. 2, 398-414.

Metz, D.E., Whitehead, R.L., and Mahshie, J.J. (1982), "Physiological correlates of the speech of the deaf: a preliminary view", in *Deafness and communication: assessment and training,* ed. R.L. Whitehead, Williams and Wilkins, Baltimore.

Microsoft Corporation (1995) "The Windows Interface: An Application Design Guide", Redmont, MA. 1995.

Miller, J.D., Engebretson, A.M., and Vemula, N.R. (1980), "Vowel normalisation: Differences between vowels spoken by children, women, and men", *J. Acoust. Soc. Am.* Suppl.1, 68, S33.

Mills, M.I. (1982), "A study of the human response to pictorial representations on Telidon", Technical report from the Department of Communications, Ottawa.

MIL-STD-1472C. (1981), "Military standard: Human engineering design criteria for military systems, equipment and facilities", Washington DC, Department of Defence.

Minifie, F.D., Thomas, J.H., and Williams, F (editors), (1973) "*Normal aspects of speech, hearing and language*", Prentice -Hall, Englewood Cliffs, New Jersey.

Monsen, R.B. (1974), "Durational aspects of vowel production in the speech of deaf children", *Journal of Speech and Hearing Research,* vol.17, pp 386-398.

Monsen, R.B. (1976a), "Normal and reduced phonological space: the production of English vowels by deaf adolescents", *Journal of Phonetics.,* vol.4, pp 29-42.

Monsen, R.B. (1976b), "The production of English stop consonants in the speech of deaf children", *J. Phon.,* vol.4, pp 29-42.

Monsen, R.B. (1976c), "Second formant transitions of selected consonant-vowel combinations in the speech of deaf and normal-hearing children", *Journal of Speech and Hearing Research.,* vol.19, pp 279-289.

Monsen, R.B. (1976d), "A taxonomic study of diphthong production in the speech of deaf children", in *Hearing and Davis: Essays honoring Hallowen Davis.,* ed. S.K. Hirsh, DH. Eldredge, I.S. Hirsh and S.R. Silverman, Washington University Press, St. Louis.

Monsen, R.B. (1978), "Toward measuring how well hearing impaired children speak", *Journal of Speech and Hearing Research,* vol.22, pp 270-288.

Monsen, R.B. (1979),"Acoustic qualities of phonation in young hearing-impaired children", *Journal of Speech and Hearing Research,* vol. 22, pp. 270-288.

Monsen, R.B., Engebretson, A.M., and Vemula, N. (1979),"Some effects of deafness on the generation of voice", *J Acoust. Soc. Amer.,* vol. 57, p. 569(A).

Mullet, K. and Sano, D. (1995), "Designing visual interfaces", Prentice Hall, 1995.

Myers, B.A. and Rosson, M.B. (1992), "Survey on user interface programming", in: Bauersfeld, P.; Bennet, J. & Linch, G. (Eds.) CHI '92 Conference Proceedings: ACM Conference on Human Factors in Computing Systems.

N

NASA (1980), "Spacelab display design and command usage guidelines", Report MSFC-PROC-711A, Huntsville, AL, George C. Marshall Space Flight Center.

Neisser, U. (1966), "Cognitive psychology", New York, Appleton-Century-Crofts..

Ney (1983), "A Dynamic Programming Approach to speech parameter estimation", IEEE Trans. System, Man and Cybernetics, March, 1993.

NeXT Corporation (1992), "NeXTSTEP User Interface Guidelines Release 3", Reading, MA, Addison-Wesley Publishing.

Nickerson, R.S., and Stevens, K.N. (1973), "Teaching Speech to the Deaf: Can A Computer Help?", *IEEE Transactions on Audio and Electroacoustics,* vol. AU-21, no.5, pp. 445-455.

Nickerson, R.S., Kalilow, D.N., and Stevens, K.N. (1976), "Computer aided speech training for the deaf", *Journal of Speech and Hearing Disorders,* 41, pp. 120-132.

Nober, E.H. (1967), "Articulation of the deaf", *Exceptional Children,* vol. 33, pp.611-621.

Nolan, Norton and Co. (1992) "Managing End-User Computing". Boston, Nolan, Norton and Co.

Noll, A.M. (1970), "Pitch determination of human speech by the harmonic product spectrum, the harmonic sum spectrum, and a maximum likelihood estimate", vol. 19 of *Symposium on Computer Processing in Communication,* pp. 779-797. Polytechnic Institute of Brooklyn Microwave Research Institute, New York, 1970.

O

OPEN Software Foundation (1993), "OSF/Motif (TM) Style Guide", Englewood Cliffs, NJ, Prentice Hall.

Osberger, M.J. (1978), "The effect of timing errors on the intelligibility of deaf children's speech", unpublished doctoral dissertation, City University of New York.

Osberger, M.J. (1981), "Fundamental frequency characteristics of the speech of the hearing-impaired", *J. Acoust. Soc. Amer.,* vol. 66, pp. 1316-1324.

Osberger, M.J., Levitt, H., and Slosberg, R. (1979), "Acoustic characteristics of correctly produced vowels in deaf children's speech", *J. Acoust. Soc. Amer.,* vol. 66, p.S13.

Osberger, M.J., Moeller, M., Kroese, J., and Lippmann, R. (1981), "Computer-assisted speech training for the hearing impaired", *Journal of the Academy of Rehabilitative Audiology,* vol. 14, pp. 145-158.

Osberger, M.J. and McGarr, N.S. (1982), "Speech production characteristics of the hearing impaired", in *Speech and Language: Advances in basic research and practice,* vol. 8, ed. N.J. Lass, pp.221-283.

P

Pakin, S.E., and Wray, P. (1982), "Designing screens for people to use easily", *Data Management*, July 1982, 36-41.

Penn, J.P. (1965), "Voice and speech patterns in the hard of hearing", *Acta Otoloaryngologica,* Suppl. 124.

Perrin, E., Borel, J., Berger-Vachon, C., and Kauffmann, I. (1994), "Phonatory signature of the deaf child", Proc. ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, Martigny, Switzerland, 5-7 April 1994, pp. 201-204.

Peterson, G.E., Barney, H.L. (1952), "Control Methods used in a study of the vowels", J.A.S.A. vol. 24, pp.175-184.

Peterson, D.E. (1979), "Screen design guidelines", *Small Systems World,* February 1979, pp.19-37.

Phillips, N., Remillard, W., Bass, S., and Pronovost, W. (1968), "Teaching of intonation to the deaf by visual pattern matching", *American Annals of the Deaf,* vol. 113, pp. 239-246.

Pickett, J.M. (1980), "The sounds of speech communication: A primer of acoustic phonetics and speech perception", University Park Press, Baltimore.

Plant, G. (1983), "The effect of a long-term hearing loss on speech production", *Speech Transmission Lab. Q. Prog. Status Rep.,* vol. 1, pp. 18-35.

Plant, G., and Hammarberg, B. (1983), "Acoustic and perceptual analysis of the speech of the deafened", *Speech Transmission Lab. Q. Prog. Status Rep.,* vol. 2-3, pp. 85-107..

Plomp, R. (1975), "Auditory analysis and timbre perception", in *Auditory Analysis and Perception of Speech,* edited by G. Fant and M.A.A. Tatham (Academic, London), pp. 7-22.

Povel, D.J., Maasen, B. (1987), "Visual information and speech acquisition of the deaf", Proc. 11th ICPLS, Tallin, vol. 1.

Povel, D.J., and Arends, N. (1991), "The visual speech apparatus: Theoretical and practical aspects", *Speech Communication,* vol. 10, pp. 59-80.

Povel, D.J., Wansink, M. (1986), "A computer-controlled vowel corrector for the hearing-impaired", *Journal of Speech and Hearing Research,* 29, pp. 99-105.

Pratt, S. Heintzelman, A.T., and Deming, S.E (1993), "The efficacy of using the IBM Speech Viewer vowel accuracy module to treat young children with hearing impairment", *Journal of Speech and Hearing Research,* vol. 36, pp. 1063-1074.

Q

R

Rabiner, L.R., Schafer, R.W., and Rader, C.M. (1969), "The Chirp z-Transform Algorithm", *IEEE Trans. Audio and Electroacoust.,* Vol. AU-17, No. 2, pp. 86-92, June 1969.

Rabiner, L.R., and Schafer, R.W. (1978), "Digital Processing of Speech Signals", Prentice-Hall, New Jersey.

Rabiner, L.R., Juang, B.H, Levinson, S.E., and Sondhi, M.M. (1985), "Recognition of Isolated Digits Using Hidden Markov Models with Continuous Mixture Densities", *AT&T Tech. J.,* Vol. 64, pp. 1211-1234.

Reilly, A.P. (1979), "Syllabic nucleus duration in the speech of hearing and deaf persons", unpublished doctoral dissertation, City University of New York.

Ringel, S., and Hammer, C. (1964), "Information assimilation from alphanumeric displays: Amount and density of information presented (Tech Report TRN141), Washington, DC, US Army Personnel Research Office.

Ryalls, J. (1989), "Comparison of two computerised speech training systems: SpeechViewer and ISTRA", *Journal of Speech-Language Pathology and Audiology,* 13 (3), pp. 53-56.

Robinson, T. (1996) , http://squid.eng.cam.ac.uk:80/~ajr/SpeechAnalysis/index.html, Cambridge Uni. Eng. Dept./Speech Vision Robotics group.

Rooney, E.J, (1990), ""Nasality in Automatic Speaker Verification", Ph.D., The University of Edinburgh.

Rooney, E.J, Hiller, S.M., Laver, J. and Jack, M.A. (1992), "Prosodic features for automated pronunciation improvement in the SPELL system", In *Proc. International Conference on Spoken Language Processing,* vol. 1, pp. 413-416, Banff, Canada.

Rossiter, D., and Howard, D.M. (1994), "Animation of larynx movement derived from an electrolaryngograph signal", *Voice* (journal of the British Voice Association), vol. 3, pp. 86-91.

Rothman, H.R. (1976), "A spectrographic investigation of consonant vowel transitions in the speech of deaf adults", *J Phonetics*, vol. 4, pp. 129-136.

Ruoss, M., Hobohm, K., and Drautzburg, M. (1988), "Visible speech. Examination of different patterns and development of a new speech visualizer", *Proceedings Speech '88: 7th FASE Symposium, Edinburgh, 22-26 August 1988,* pp. 187-194.

S

Sacks, Oliver (1989), "Seeing voices", University of California Press, 1989.

Sanders, A.F. (1970), "Some aspects of the selective process in the functional visual field", *Ergonomics,* 13 (1), 101-118.

Schroeder, M.R. (1968), "Period histogram and product spectrum: New methods for fundamental frequency measurement", *J. Acoust. Soc. Amer,* 43 (4), pp. 829-834.

Schroeder, M.R., Atal, B.S., and Hall, J.L. (1979), "Objective measure of certain speech signal degradation based on masking properties of human auditory perception", in *Frontiers of Speech Communication Research,* edited by B. Lindblom and S. Ohman (Academic, London), pp.217-229.

Schwartz, D.R. (1986), "Formatting effects on the use of computer-generated alphanumeric displays: The moderating effects of task characteristics", unpublished doctoral dissertation, Rice University.

Schwartz, D.R., and Howell, W.C. (1985), "Optional stopping performance under graphic and numeric CRT formatting", *Human Factors*, 27, 433-444.

Seaver, E.G., Andrews, J.R., and Granata, J.J. (1980), "A radiographic investigation of velar positions in hearing impaired young adults", *Journal of Communication Disorders,* vol. 3, pp. 239-247.

Secrest, B.G., and Doddington, G.R. (1983), "An integrated pitch tracker algorithm for speech systems", in: *Proc. IEEE ICAASP-83,* pp. 1352-1355, Boston.

Sharf, B. (1970), "Critical bands", in *Foundations of Modern Auditory Theory,* edited by J. V. Tobias (Academic, New York), Vol. 1, pp. 157-200.

Shiffrin, R.M., and Schneider, W. (1977), "Controlled and automatic human information processing II: Perceptual learning, automatic attending, and a general theory", *Psychological Review,* 84, pp. 127-190.

Shirai, K. and Honda, M. (1980), "Estimation of articulatory motion from speech waves and its application for automatic recognition", in *Spoken Language Generation and Understanding,* (ed. J. C. Simon), Reidel, Dordrecht, Holland, pp. 87-99.

Smith, S. (1951), "Vocalisation and Added Nasal Resonance", *Folia Phoniatrica,* Vol. 3, pp. 165-169.

Smith, S. (1954), "Remarks on the Physiology of the Vibrations of the Vocal Cords", *Folia Phoniatrica,* Vol. 6, pp. 166-178.

Smith, C. (1975), "Interjected sounds in deaf children's speech", *J Communication Disorders,* vol. 8, pp. 123-128.

Smith, C. (1975b), "Residual hearing and speech production in deaf children", *Journal of Speech and Hearing Research,* vol. 18, pp. 795-811.

Smith, S.L., and Mosier, J.N. (1986), "Guidelines for designing user interface software", Hanscom Air Force Base, MA, USAF Electronic Systems Division.

Smith, J.D. (1984), "Overall Similarity in Adults' Classification: The Child in All of Us", *Journal of Experimental Psychology: General,* 1984, Vol. 113, No. 1, 137-159.

Sonesson, B. (1960), "On the Anatomy and Vibratory Pattern of Human Vocal Folds", Thesis, Lund..

Spaai, G.W.G. (1983), "Teaching intonation to the deaf through visual displays", In B. Elsendoorn and F. Coninx (Eds.), *Proceedings of the NATO-ARW conference on 'Interactive Learning Technology for the Deaf', ARW-series F,* pp. 151-163, Springer-Verlag, Berlin.

Stewart, T.F.M. (1976), "Displays and the software interface", *Applied Ergonomics,* 7.3, 137-146.

Stevens, K., Boothroyd, A., and Rollins, A. (1976), "Assessment of nasalization in the speech of deaf children", *Journal of Speech and Hearing Research,* vol. 19, pp. 393-416.

Stevens, K. (1977), "Physics of laryngeal behaviour and larynx models", *Phonetica,* 34, pp. 254-279.

Stevens, K., Nickerson, R., and Rollins, A. (1978), "On describing the suprasegmental properties of the speech of deaf children", in *Advances in prosthetic devices for the deaf: a technical workshop,* ed. D. McPherson and M. Davids, National Technical Institute for the Deaf, Rochester, New York, pp. 134-155.

Stevens, K., Nickerson, R., and Rollins, A. (1983), "Suprasegmental and posture aspects of speech production and their effect on articulatory skills and intelligibility", in *Speech of the hearing impaired: research, training and personnel preparation,* ed. I. Hochberg, H. Levitt and M.N. Osberger, University Park Press, Baltimore, MD.

Streveler, D.J., and Wasserman, A.I. (1984), "Quantitative measures of the spatial properties of screen designs", *Proceedings of INTERACT '84 Conference on Human-Computer Interaction,* London, England, Sept.1984.

Strevens, P. (1960), "Spectra of fricative noise in human speech", *Lang. and Speech,* vol. 3, pag. 32-49.

Stoker, R., Fitzgerald, M., and Gruenwald, A. (1987), "Review of video voice speech training system", *The Volta Review,* 89 (3), pp. 171-173.

Sydral, A.K., and Gopal, H.S. (1986),"A perceptual model of vowel recognition based on the auditory representation of American English vowels", *J. Acoust. Soc. Am.,* 79 (4), April 1986.

T

Tartter, V., Chute, P., and Hellman, S. (1989), "The speech of a postlingually deafened teenager during the first year of use of a multichannel cochlear implant", *J. Acoust. Soc. Amer.,* vol. 86, pp. 2113-2121.

Traunmuller, H. (1981), "Perceptual dimension of openness in vowels", *Journal of the Acoustical Society of America,* 69, 1465-1475.

Treisman, A. (1986), "Features and Objects in Visual Processing", *Scientific American,* n. 255, Nov 1986.

Tullis, T.S. (1983), "The formatting of alphanumeric displays: A review and analysis", *Human Factors,* 25, 657-682.

Tullis, T.S. (1984), "Predicting the Usability of Alphanumeric Displays", PhD Dissertation, Rice University, Lawrence, Kansas.

Tullis, T.S. (1988), "Screen design", in: *Handbook of Human-Computer Interaction,* M.Helander (ed.), Elsevier Science Publishers B.V. (North Holland), 1988.

U

Ui-Design (1996) Industrial Design Engineering World Wide Web of Delft University of Technology (http://www.io.tudelft.nl/uidesign). The Netherlands.

V

Vieira, M.N., McInnes, F.R., and Jack, M.A. (1985), "Effects of the EGG baseline fluctuation on the F0 estimation in pathological voices", submitted to the *J. Acoust. Soc. Am.*

Vieira, M.N., McInnes, F.R., and Jack, M.A. (1986), "Time-domain F0 estimation in pathological connected speech", *The Postgraduate Journal of the Department of Electrical Engineering,* Vol. 2, Jan. 1996, University of Edinburgh.

W

Waldstein, R. (1990), "Effects of postlingual deafness on speech production: implications for the role of auditory feedback", *J. Acoust. Soc. Amer.,* vol. 88, pp. 2099-2114.

Watanabe, A., Ueda, Y., and Shigenaga, A. (1985), "Color display system for connected speech to be used for the hearing impaired", *IEEE Transactions on Acoustic, Speech, and Signal Processing,* Vol ASSP-33, 1, 164-173.

Watanabe, A., Tokunaga, S., and Tomishige, S. (1995), "Speech visualization by extracting features with neural networks", *15th International Congress on Acoustic,* Trondheim, Norway 26 - 30 June 1995.

Watson, C.S. and Kewley-Port, D. (1989), "Advances in Computer-Based Speech Training: Aids for the Profoundly Hearing Impaired", *The Volta Review*, Vol. 91 No. 5.

Watson, C.S., Reed, D.J., Kewley-Port, D., Maki, D. (1989), "The Indiana Speech Training Aid (ISTRA) I: Comparisons between human and computer-based evaluation of speech quality", *Journal of Speech and Hearing Research,* Vol. 32, pp. 245-251, June 1989.

Watson, C.I., Kennedy, W.K., Bates, R.H.T. (1990), "Towards a computer based speech therapy aid", *Proceedings 3rd AICSST,* Melbourne.

Wempe, T.G., and Lunen, M. Van (1991), "The IBM SpeechViewer", *Proceedings of the Institute of Phonetic Sciences Amsterdam,* 15, pp. 121-129.

Whitehead, R. (1983), "Some respiratory and aerodynamic patterns in the speech of the hearing-impaired", in *Speech of the hearing impaired: research, training and personnel preparation,* ed. I. Hochberg, H. Levitt and M.N. Osberger, University Park Press, Baltimore, MD.

Whitehead, R. and Barefoot, S. (1980), "Some aerodynamic characteristics of plosive consonants produced by hearing-impaired speakers", *American Annals of the Deaf,* vol. 125, pp.366-373.

Whitehead, R. and Emanuel, F.W. (1974), "Some spectrographic and perceptual features of vocal fry", *Journal of Communication Disorders,* vol. 7, pp.305-319.

Williges, B.H., and Williges, R.C. (1981), "User considerations in computer-based information systems" (Technical Report CSIE-81-2), Blacksburg, VA, Virginia Polytechnic Institute and State University.

Wirz, S. (1987), "Vocal characteristic s of hearing-impaired people", unpublished doctoral thesis, University of Edinburgh. (Source: DS1)

Wirz, S., Subtelny, J., amd Whitehead, R. (1979), "A perceptual and spectrographic study of tense voice in normal hearing and deaf subjects", *Fol Phon,* vol. 33, pp. 23-36.

Wolf, C.E. (1986), "BNA 'HN' command display: Results of user evaluation", unpublished tech. report, Unisys Corporation, 19 Morgan, Irvine, CA.

Wrench, A.A., Jackson, M.S., Soutar, D.S., Robertson, A.G., and MacKenzie Beck, J. (1995), "Evaluation of a system for segmental speech quality assessment: voiceless fricatives", *Proceedings EuroSpeech 95,* Madrid, pp. 1879-1882.

Wrench, A.A., McIntosh, A.D. and Hardcastle, W.J. (1996), "Optopalatograph (OPG): a new apparatus for speech production analysis", *Proceedings ICSLP 96,Philadelphia*, pp. 1589-1592.

X

Y

Yamada, Y., Murata, N., and Oka, T. (1988), "A new speech training system for profoundly deaf children", *Journal of the Acoustical Society of America,* 84 (Suppl. 1), S43.

Yamada, Y., Murata, N. (1991), "Computer Integrated Speech Training Aid", *Proc. International Symposium on Speech and Hearing Sciences*, Osaka, July 1991.

Yorchak, J.P., Allison, J.E., and Dodd, V.S. (1984), "A new tilt on computer generated space situation displays", *Proceedings of the Human Factors Society 28th Annual Meeting,* 894-898, Santa Monica, CA, Human Factors Society.

Z

Zimmerman, G., and Rettaliata, P. (1981), "Articulatory patterns of an adventitiously deaf speaker: implications for the role of auditory information in speech production", *Journal of Speech and Hearing Research,* vol. 24, pp. 169-178.

Zwicher, E. (1961), "Subdivision of the audible frequency range into critical bands (Frequenzgruppen)", *Journal of the Acoustical Society of America.* 33, 248.